**Universidade de Aveiro**
**2009**
Departamento de Electrónica,
Telecomunicações e Informática

**João Manuel Castro Fernandes**

# GeneBrowser – Sistema de recuperação de dados biológicos

**Universidade de Aveiro**
**2009**
Departamento de Electrónica,
Telecomunicações e Informática

**João Manuel Castro Fernandes**

# GeneBrowser – Sistema de Recuperação de Dados Biológicos

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Electrónica e Telecomunicações, realizada sob a orientação científica do Professor Doutor José Luís Guimarães Oliveira, Professor Associado do Departamento de Electrónica, Telecomunicações e Informática da Universidade de Aveiro

Para a minha família por todo o apoio e compreensão e a todos aqueles que sempre me apoiaram.

**o júri**

presidente

Doutor Joaquim Arnaldo Carvalho Martins
Professor Catedrático da Universidade de Aveiro

Doutor Rui Pedro Sanches de Castro Lopes
Professor Coordenador da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança

Doutor José Luís Oliveira
Professor Associado da Universidade de Aveiro

**agradecimentos**

Ao Professor José Luís Oliveira e ao Joel Arrais pelos conselhos, críticas e sugestões durante o desenvolvimento deste trabalho; à minha família e aos meus amigos por todo o incentivo e apoio que me deram ao longo do ano.

Gostaria ainda de agradecer ao Luís Santos pelo seu contributo na área de biologia, ao João Pereira pelo trabalho desenvolvido no GeNS e ao resto do grupo de Bioinformática pelo seu companheirismo.

Por fim, a todos aqueles cuja memória me possa ter falhado.

**palavras-chave**

Web2.0, javascript, xml, ajax, css, serviços web, html, conhecimento, json, Integração de dados, base de dados biológicas, extracção de dados, serviços

**resumo**

O uso de técnicas computacionais tem sido um aspecto decisivo no avanço da biologia molecular e genética. Toda a pesquisa moderna da biologia molecular não seria possível sem a computação. Da sinergia existente entre estas duas ciências surgiu um novo campo do conhecimento - a bioinformática. A sequenciação do genoma humano e de outros organismos criou uma enorme quantidade de dados, o que levou à necessidade de criar novas aplicações para processar e extrair informação destes dados.

Esta dissertação teve como objectivo, planear e desenvolver uma solução computacional que permitisse uma melhor interpretação biológica de um conjunto de genes. O sistema construído, *GeneBrowser*, agrega diversos serviços e fontes dados públicos, e utiliza métodos estatísticos para evidenciar o enriquecimento funcional das classes que descrevem os genes e os seus produtos.

**keywords**

Web2.0, javascript, xml, ajax, css, web services, html, services, knowledge, json, Data integration, biological databases, data extraction, service

**abstract**

The use of computational techniques is a decisive aspect in the advances of molecular biology and genetics. All the modern research in molecular biology was not possible without the computation techniques. With the synergy between the two sciences a new field of knowledge appeared – bioinformatics. With the sequencing of the human genome and other organisms, a huge amount of data was created, this brought the need of new applications for process and extract knowledge from this data.
This work had as objectives, to plan and develop a computational solution that provides a better biological interpretation of a set of genes. The system built, *GeneBrowser*, adds several services and public data sources, and uses statistical methods to demonstrate the functional enrichment of the classes that describe the genes and their products.

# Índice

# Índice de Figuras

# Índice de Tabelas

# Acrónimos

| | |
|---|---|
| ***AJAX*** | *Assynchronous Javascript and XML* |
| ***API*** | *Application Programming Interface* |
| ***CSS*** | *Cascade Style Sheet* |
| ***DB*** | *Database* |
| ***DBMS*** | *DataBase Management System* |
| ***DOM*** | *Document Object Model* |
| ***FAQ*** | *Frequently Asked Questions* |
| ***FTP*** | *File Transfer Protocol* |
| ***HTML*** | *HyperText Markup Language* |
| ***HTTP*** | *HyperText Transfer Protocol* |
| ***JSON*** | *JavaScript Object Notation* |
| ***PDA*** | *Personal Digital Assitant* |
| ***SGBD*** | *Sistema de Gestão de Bases de Dados* |
| ***SOA*** | *Service Oriented Architecture* |
| ***SOAP*** | *Simple Object Access Protocal* |
| ***UDDI*** | *Universal Description, Discovery and Integration* |
| ***URL*** | *Uniform Resource Locator* |
| ***WSDL*** | *Web Services Description Language* |
| ***XML*** | *eXtensible Markup Language* |
| ***Xpath*** | *XML Path Language* |
| ***XSL*** | *eXtensible Stylesheet Language* |

# Capítulo 1 – Introdução

Da união das duas ciências que mais evoluíram nos últimos anos, a biologia molecular e a informática, surgiu um novo campo do conhecimento – a bioinformática.

Na realidade, devido aos avanços da ciência, o biólogo de hoje deixou de ter só o microscópico como principal ferramenta de trabalho e passou a poder contar também com o computador, o que gerou uma demanda de aplicações informáticas capazes de satisfazer as suas necessidades [1]. A bioinformática é uma nova área da ciência que analisa a informação extraída dos projectos de investigação nas diversas áreas da biologia, genómica e proteinómica.

Como campo vasto que é, a bioinformática levanta muitos novos paradigmas, mas é também fonte de um conjunto de aplicações que apresentam soluções para diversos problemas, incluindo o do estudo de expressão dos genes, actualmente bastante explorado através da tecnologia de microarrays de DNA (ferramenta analítica bastante versátil, podendo ser adaptados a aplicações que vão desde o diagnóstico de doenças genéticas até à monitorização do ambiente, sendo a sua principal vantagem a de permitir uma larga escala de amostragem).

## 1.1. Enquadramento

Numa experiência de microarray existe um conjunto de passos a serem efectuados de modo a atingirmos o resultado esperado, em que o trabalho laboratorial é apenas o primeiro de um conjunto de tarefas encadeadas que inclui o trabalho computacional e análise de resultados. A forma clássica de se fazer o trabalho computacional é começar pela aquisição de dados, à qual se segue uma análise de qualidade e pré-processamento (Figura 1.1). Durante a análise exploratória, vários métodos estatísticos (*Fold Change*, ANOVA e *Support Vector Machines*) podem ser usados para obtermos a lista de genes diferencialmente expressos.

Neste momento, no laboratório de biologia molecular da Universidade de Aveiro, já se encontra em funcionamento um sistema informático que possibilita a gestão de todo o fluxo laboratorial relacionado com experiências de *microarrays*. Este sistema, designado por *Mind* (*Microarray Information Database*) está disponível em http://bioinformatics.ua.pt/mind, permite realizar o

registo de todo o processo laboratorial, assim como a visualização e o respectivo tratamento estatístico dos dados adquiridos.

O *Mind* permite o suporte de todas as etapas até à análise exploratória (Figura 1.1), ficando apenas de fora a análise de resultados. Neste contexto, percebesse que que um dos principal objectivo do *Mind*, para além das etapas de anotação da experiência, é obter a lista de genes diferencialmente expressos, dados estes que serão o principal ponto de entrada do *GeneBrowser*.

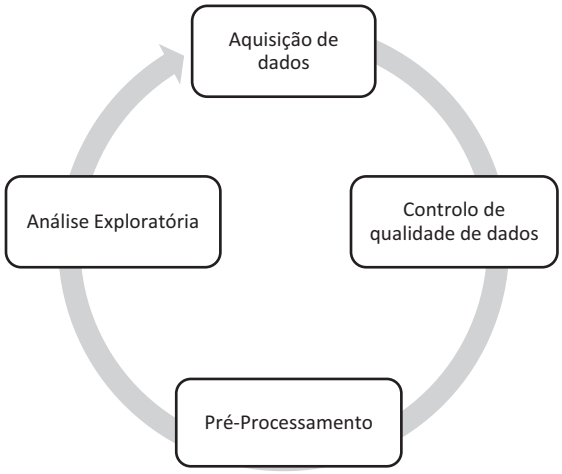



**Figura 1.1 Etapas realizadas no trabalho computacional realizado pelo *Mind*.**

Uma das possibilidades para a análise de resultados consiste em representar os genes segundo uma categoria e função biológica (Figura 1.2), uma vez que é aceite que os genes com uma expressão similar tendem a ter regras biológicas similares. Programas como *Onto-Express* [2], *FatiGO* [3] e *GoMiner* [4] podem ser considerados como ferramentas de referência para uma análise funcional e ontológica. Outro conjunto de ferramentas completamente diferentes pode ser usado para encontrar relações entre os genes em termos de vias de sinalização em que os genes estão envolvidos. Ferramentas como *GenMAPP* [5] e *PathwayExplorer* [6] podem ser usadas nesta tarefa. A complexidade dos fenómenos biológicos envolvidos requer o uso de uma aproximação mais sistemática, que combina a informação fornecida por esta ferramenta com as descrições dos genes e a informação específica disponibilizada na literatura científica, a fim de extrair o conhecimento consistente dos dados.

Para esta finalidade, foi desenvolvido previamente um (portal/serviço) *Web* chamado *GeneBrowser* [7], que combina a informação de diversas fontes de dados e métodos de visualização. Esta ferramenta tem como ponto de entrada, uma lista de genes e um organismo. De seguida integram-

se os dados correspondentes a diversas fontes públicas, tais como o *Gene Ontology* [8], *KEGG* [9], *PubMed* [10] e extrai-se a informação relevante para cada gene. Estes dados são processados e combinados, dando ao utilizador a oportunidade de explorar os resultados, usando diversos métodos de visualização. Além disso, o sistema fornece as ligações directas para visualizar os dados na fonte e contexto original onde os dados complementares são fornecidos.

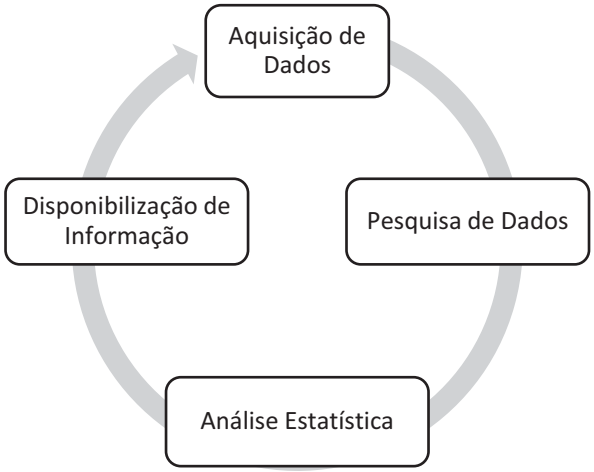



**Figura 1.2 Etapas a realizar na análise de resultados.**

## 1.2. Motivação

A motivação que dirigiu este trabalho foi construir um novo sistema que sucedesse ao *GeneBrowser*. Este sitema deverá combinar as vantagens já existentes no *GeneBrowser* com vários outros sistemas como o *Entrez* [11], *Uniprot* [12], *InterPro* [13], *Pfam* [14], *GeneCards* [15] *GoMiner* [16], *Onto-Express* [2] ou o *PathwayExplorer* [6], aumentando assim, não só o número de fontes de dados integradas, mas contendo também as características actuais de outros softwares usados para a extracção do conhecimento das séries de dados microarray de DNA.

## 1.3. Objectivo

O principal objectivo deste trabalho consistiu no desenvolvimento de uma ferramenta que possibilitasse a anotação e análise de um conjunto de genes provenientes de uma experiência de expressão genética. A ferramenta deveria ter a capacidade de aceder a bases de dados externas e de obter dados que permitissem documentar cada um dos genes. De seguida, deveria realizar a distribuição e enriquecimento funcional dos genes anotados de acordo com diferentes critérios. A anotação de genes devia conter: ortologias, vias metabólicas, homologias, localização cromossómica, expressão génica e informação diversa para permitir identificar e caracterizar cada gene.

A informação a colocar disponível visualmente seria: lista de genes anotados, imagem da via metabólica com o gene seleccionado, informação adicional sobre o gene seleccionado e link directo para base de dados externas (KEGG, GO, SGD), gráfico com distribuição e enriquecimento de genes por via metabólica, ontologia, homologia, cromossoma, árvore de ontologia enriquecida com p-value, outras experiências de expressão génica onde esses genes estiveram envolvidos e os artigos relacionados com os genes disponíveis através do NCBI-*Pubmed.*

Devido ao interesse em utilizar outras bases de dados, o sistema deveria ser desenvolvido de forma a suportar futuras actualizações. Deste modo o acesso às bases de dados externas seria realizado preferencialmente através de *Web Services*. No entanto, caso esta opção não fosse viável deveria estar contemplada a possibilidade de instalar localmente um mirror.

## 1.4. Actividades Desenvolvidas

Para atingir os objectivos desta dissertação foi seguido um plano de trabalhos proposto inicialmente.

Primeiro trabalhou-se na especificação de requisitos técnicos e funcionais, desde conhecer as fontes de dados até à análise de sistemas similares. Depois passou-se à fase de desenvolvimento que foi dividida em várias etapas discriminadas na Tabela 1.1 e Figura 1.3. Este trabalho foi concluído fazendo o sistema passar por uma fase de testes e com a fase final de escrita da tese.

**Tabela 1.1 Cronograma das etapas de execução do trabalho de mestrado.**

| # | Nome | Duração em Semanas |
|---|---|---|
| 1 | GeneBrowser 2.0 | 46 |
| 2 | Especificação de Requisitos | 6 |
| 3 | Desenvolvimento | 36 |
| 4 | Primeira avaliação e correcção de erros | 4 |
| 5 | Data Processing | 3 |
| 6 | Bibliography | 2 |
| 7 | Segunda avaliação e correcção de erros | 2 |
| 8 | Tese | 4 |

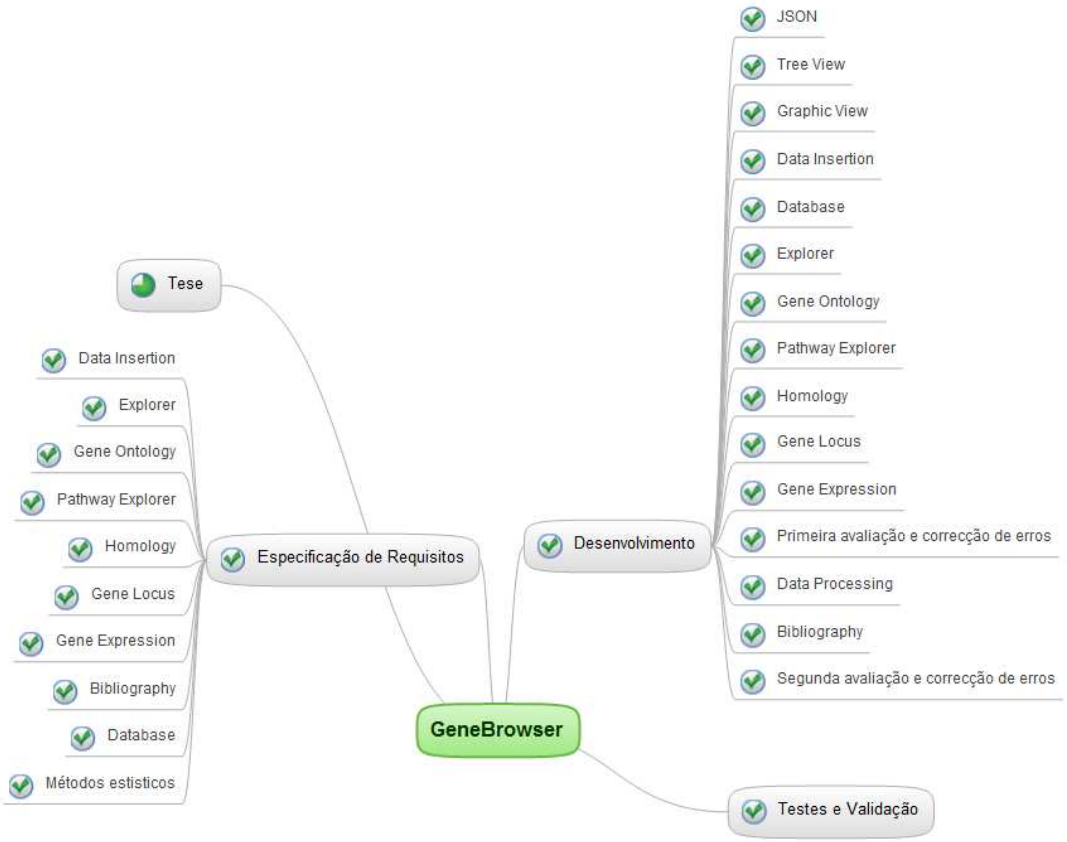



**Figura 1.3 Mapa das etapas executadas no trabalho de mestrado.**

## 1.5. Organização da Dissertação

Este documento encontra-se dividido em sete capítulos.

O capítulo 2 apresenta a problemática que motivou a construção da aplicação, uma breve descrição de fontes de dados a utilizar na ferramenta, bem como uma análise de algumas ferramentas já existentes e os problemas encontrados.

No capítulo 3 observa-se um estudo detalhado das tarefas passíveis de serem implementadas no âmbito deste trabalho e é apresentada uma proposta de arquitectura para a aplicação.

O capítulo 4 apresenta a estratégia de desenvolvimento da arquitectura.

No capítulo 5 são apresentados os testes e a validação dos casos possíveis de utilização da ferramenta.

Finalmente, o capítulo 6 apresenta as principais conclusões do trabalho desenvolvido e algumas sugestões para trabalhos futuros.

# Capítulo 2 - Integração de Dados Biológicos

## 2.1. Perspectiva Histórica da Bioinformática

A bioinformática é uma área recente que recebe contributos de diversas áreas como a computação, biologia molecular e a estatística. Apesar de existirem muitas definições de bioinformática, esta pode definir-se como sendo a pesquisa e desenvolvimento de ferramentas computacionais, matemáticas e estatísticas para a resolução de problemas na Biologia [17]. Pode ainda definir-se como uma combinação de Ciência da Computação com Tecnologia de Informação e genética que tem por objectivo determinar e analisar informação genética [18].

Vendo de um ponto de vista histórico o computador apareceu em meados de 1940 e a descoberta da hélice dupla, por Watson e Crick foi feita em 1953. Esta mostrou que a informação genética também é armazenada num formato digital, mas difere do alfabeto binário dos computadores. Os dados genéticos são armazenados com um alfabeto quaternário A, C, G e T. Mais tarde descobre-se que o funcionamento dos genes também é digital, até certo ponto, pois os genes podem ser "ligados" ou "desligados". Estas observações seriam suficientes para prever, na década de 1950, que um dia informática e biologia molecular se iriam juntar, originando uma nova área de conhecimento. Algumas pessoas consideram que a bioinformática passou a ser reconhecida pelo mundo científico por volta de 1995, ano em que foi publicado o primeiro genoma de uma bactéria. Esta demora no aparecimento da bioinformática é fácil de se explicar, do lado da biologia apesar da estrutura do DNA ter sido descoberta em 1953, a informação nela contida não podia ser "lida". Foi como se tivéssemos descoberto o alfabeto utilizado para escrever "o livro da vida", mas as "palavras" desse livro estavam com letrinhas tão pequenas que não conseguiam ser lidas. Foi preciso esperar até fins da década de 1980 para que aparecesse uma "lente de aumento" suficientemente boa que permitisse a leitura dessas letrinhas em grande quantidade, Microscópio Electrónico. Quanto à computação foi também preciso um amadurecimento. O computador torna-se capaz de armazenar cada vez mais informação, e processar essa de uma forma cada vez mais rápida, a um custo cada vez menor. Se a sequenciação automático do DNA tivesse amadurecido mais cedo, digamos com 20 anos de antecedência, não haveria computadores com poder suficiente para processar e guardar os dados gerados. Na década de 70 a unidade básica de armazenamento de informação era o kilobyte (1000 bytes). Um computador de grande porte daquela época tinha

alguns kbytes de memória. Com tal memória um computador desses não seria capaz de processar nem sequer o genoma de um vírus, que pode chegar a 20 kilobases.

De um ponto de vista superficial, podemos ver que a biologia molecular e a informática tiveram uma evolução sincronizada. Em 1995, quando os computadores já tinham poder de processamento suficiente e já se conseguia ler a sequência do genoma, surge a bioinformática, com a missão de nos ajudar a entender a história que está escrita nesse livro da vida (genoma).

### 2.1.1. Importância da bioinformática

A bioinformática actua em duas classes de problemas: a primeira é chamada de problema biotecnológico e a segunda classe de problemas e diz respeito à natureza da biologia molecular. Falando da primeira classe, existem dezenas ou centenas de problemas na bioinformática, cada um deles motivado por uma tecnologia em particular. Mas problemas deste tipo também existem noutras ciências, certamente os telescópios modernos geram grandes quantidades de dados de forma e formato que exigem programas de computador sofisticados para recolha e interpretação; e quando mudam os telescópios mudam os programas. A segunda classe de problemas que têm um interesse que vai além de tecnologias específicas, diz respeito à natureza da biologia molecular. São esses problemas que tornam especial a bioinformática, esses problemas são basicamente de dois tipos: Primeiro, temos a interpretação do DNA como uma linguagem, a linguagem dos genes e o segundo tipo de problema é o de entender os efeitos da informação genética.

## 2.2. Fontes de Dados

Em 2006, o número de bases de dados biológicas, públicas e comerciais existentes já ultrapassava as 1000 [19]. Estas bases de dados contêm, principalmente, informação do genoma, taxonómica e proteica. Os dados são sequências de nucleótidos, genes, aminoácidos e proteínas. Além disso contém informação sobre a função, estrutura, localização no cromossoma, efeitos clínicos das mutações, podendo também ser encontradas similaridades de sequências biológicas entre muitas outras coisas.

As bases de dados biológicas contêm um vasto conjunto de dados, facilmente acessível, organizado e permanente, e disponibilizam, frequentemente, ferramentas computacionais desenvolvidas para

actualizar, pesquisar e recolher dados armazenados no sistema. De modo a dar resposta às exigências dos investigadores que beneficiam da informação armazenada nessas bases de dados, estas devem permitir um acesso livre e fácil à informação (através da Internet) e disponibilizar métodos para extrair só a informação necessária para dar resposta a um problema biológico específico.

Podemos encontrar uma lista de base de dados na *NAR Database Categories List*. Esta lista foi realizada "*National Center for Biotechnology Information, National Library of Medicine National Institutes of Health Bethesda, MD 20894, USA*", e tem sido actualizada todos os anos.

A colecção de bases de dados de biologia molecular da NAR é um recurso público que se encontra online e contém *links* para um vasto número de bases de dados. Na actualização de 2009 o número de bases de dados de biologia molecular era 1170 [20]. Também tem uma pequena descrição de cada base de dados. A cada base de dados foi dado um número de identificação que não muda mesmo que a base de dados modifique a localização. A lista de bases de dados completa juntamente com os seus resumos pode ser encontrada no *website*: http://www.oxfordjournals.org/nar/database/a/.

Uma parte importante do trabalho a desenvolver, consiste na integração de dados biológicos. Apesar do extenso número de bases de dados, algumas destas integram dados de mais que uma fonte, tornando-se estas as mais relevantes para responder aos requisitos do sistema.

Em seguida serão apresentadas as principais instituições que trabalham na área e que contêm as fontes de dados que se tornam mais relevantes no desenvolvimento deste trabalho.

### 2.2.1. EBI

O EBI (*Instituto Europeu de Bioinformática*) [21] é parte do EMBL (*Laboratório Europeu de Biologia Molecular*), é uma instituição que contém um grande número de fontes de dados. O EBI efectua pesquisa académica em biologia molecular, genética, medicina e agricultura, e efectua pesquisa industrial em agricultura, biotecnologia, química e farmácia. Toda esta pesquisa é efectuada construindo, mantendo e disponibilizando publicamente bases de dados e serviços com informação relevante para cada um dos campos referidos.

Segundo a instituição (EBI), o seu objectivo é garantir o crescimento de informação na biologia molecular e pesquisa do genoma e colocar esta informação em lugares de acesso público, de modo a propiciar o progresso científico. Alguns dos recursos públicos mais importantes desenvolvidos pelo EBI são: *Uniprot* [12], *Ensembl* [22], *Interprot* [13] e o *ArrayExpress* [23].

O *UniProt* (*Universal Protein Resource*) (Figura 2.1) foi criado da junção de informação do *Swiss-Prot*, *TrEMBL*, e *PIR*. É um repositório de proteínas, contendo ainda a informação sobre genes, descrição de proteínas e genes, sequência, homologias, ontologias, para além de muita outra informação e acesso para bases de dados externas.

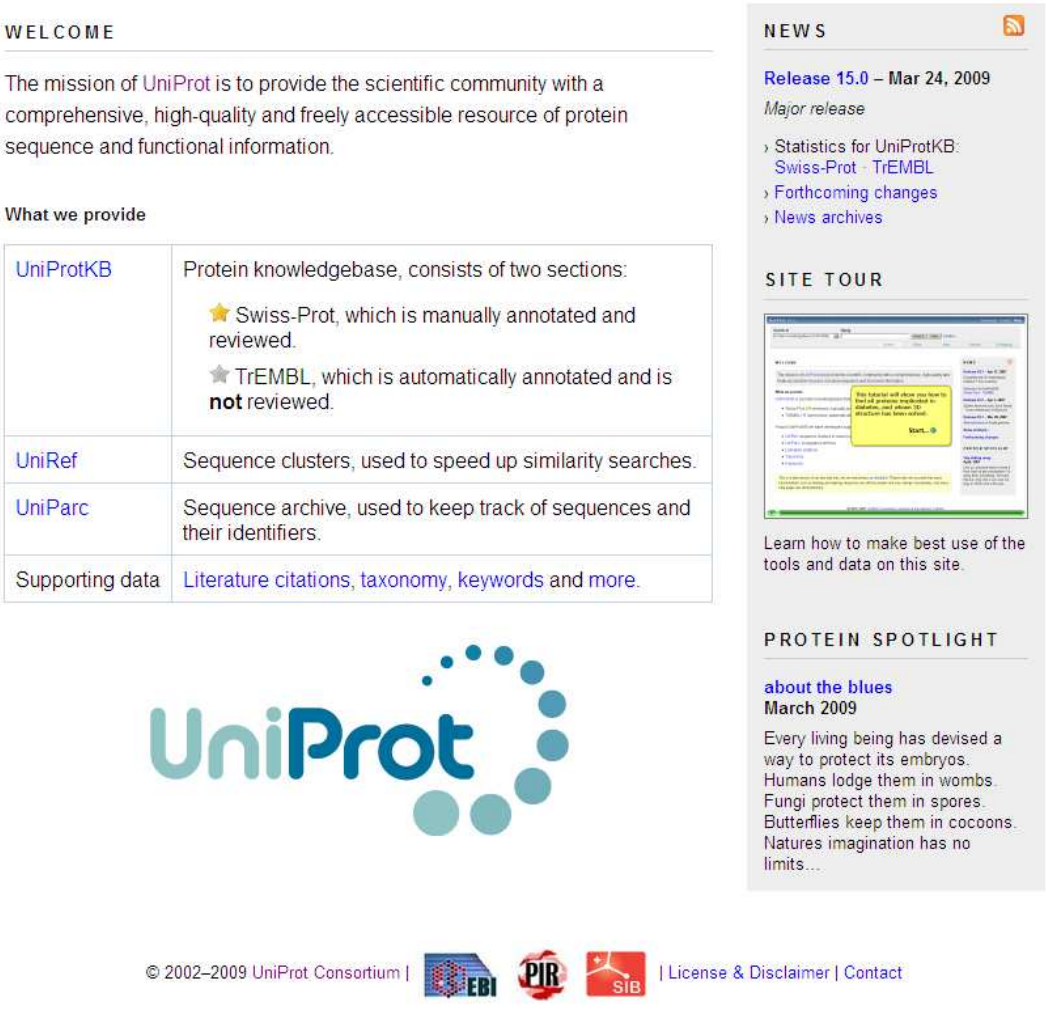



**Figura 2.1 Página inicial do Uniprot [http://www.uniprot.org/].**

A base de dados do *UniProt* é constituída por *UniProtKB (UniProt Knowledgebase)*, o *UniRef (UniProt Reference Clusters)*, o *UniMes(UniProt Metagenomic and Environmental Sequences)* e

os *Uniparc* (*UniProt Archive*) (Figura 2.2). Desta junção e do trabalho desenvolvido posteriormente, o Uniprot é hoje uma das maiores bases de dados de proteínas.

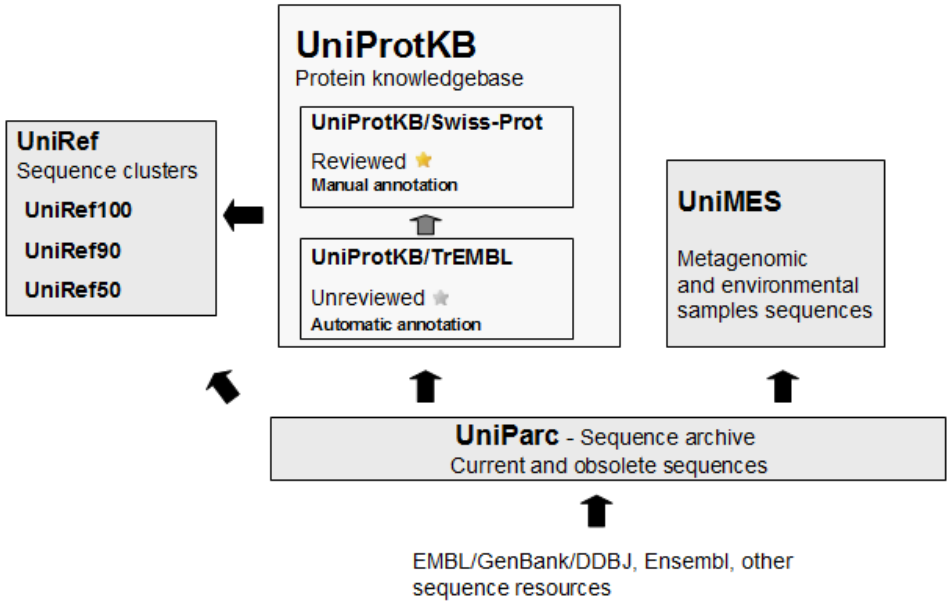



**Figura 2.2 Estrutura interna do Uniprot [http://www.uniprot.org/].**

O *Ensembl* (Figura 2.3) foi criado com o objectivo de efectuar anotação automática do genoma humano, de integrar esta anotação com outros dados biológicos disponíveis e de colocar esta informação disponível na *Internet*. Desde o seu lançamento, em Julho de 2000, muitos outros genomas foram adicionados ao *Ensembl* e o conjunto de dados disponíveis também tem expandido para incluir comparação, regulação e a variação do genoma.

O *InterPro* (Figura 2.4) integra juntamente modelos preditivos ou "conjuntos" que representam domínios, famílias e funções de proteínas, e foi contruído partindo de diversas fontes de dados: *Gene3D* [24], *Panther* [25], *Pfam* [14], *PIRSF* [26], *Prints* [27], *ProDom* [28], *Prosite* [29], *SMART* [30], *SuperFamily e TIGRFAMs* [13].

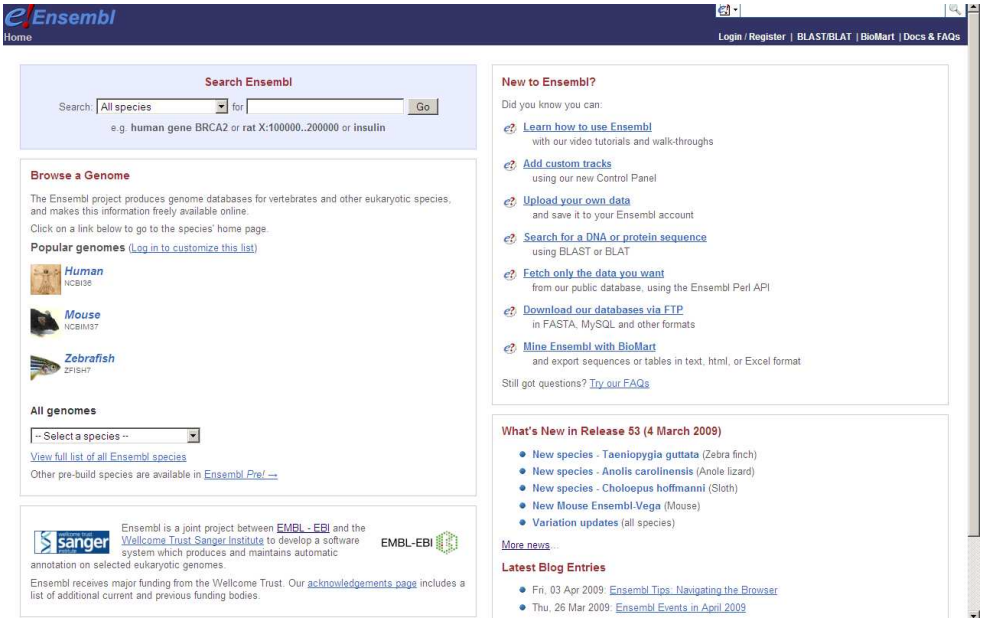



**Figura 2.3 Página inicial do Ensembl [http://www.ensembl.org/index.html].**

A integração destas fontes de dados é realizada manualmente e aproximadamente metade do total de mais ou menos 58 000 tem correspondência directa disponível no *InterPro*. O interface *Web* foi expandido e agora também contém *links* para o ADAN, base de dados de interacção proteína-proteína e para *SPICE* e *Dasty*. A última versão lançada (v18.0) tem a cobertura de 79,8% do *UniProtKB* (v14.1) e contém 16 549 entradas. Os dados do *InterPro* podem ser acedidos através do endereço *Web* (http://www.ebi.ac.uk/interpro/), através de *Web services*, por *download* dos arquivos por FTP ou usando o *software* e pesquisa *InterProScan* (http://www.ebi.ac.uk/Tools/InterProScan/) [31-33].

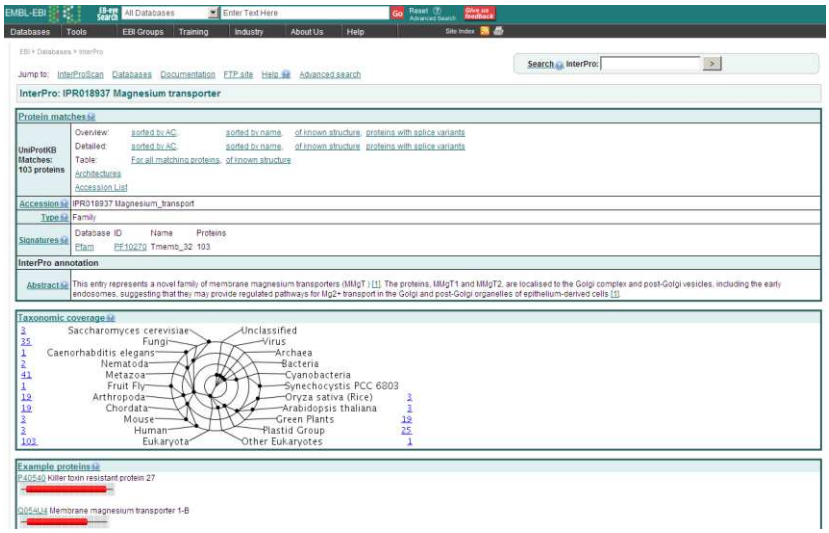



**Figura 2.4 Selecção de uma homologia no InterPro [http://www.ebi.ac.uk/interpro/].**

O *ArrayExpress* (Figura 2.5) é uma base de dados pública que guarda dados de *microarray* anotados, incluindo expressão génica, hibridação do genoma comparativa (CGH) e *Chips* de imuno-precipitação de cromatina *"chromatin-immunoprecipitation"*. O *ArrayExpress* contém mais de 50 000 hibridações, mais de 1 500 000 perfis de expressão individuais [34], mais de 6000 experiências e aproximadamente 200 000 medições [35]. O *ArrayExpress* é constituído por três grandes componentes principais a partir dos quais conseguimos obter dados de experiências anteriores: o arquivo de experiências, a *Data Warehouse* e o *ArrayExpress Atlas*.

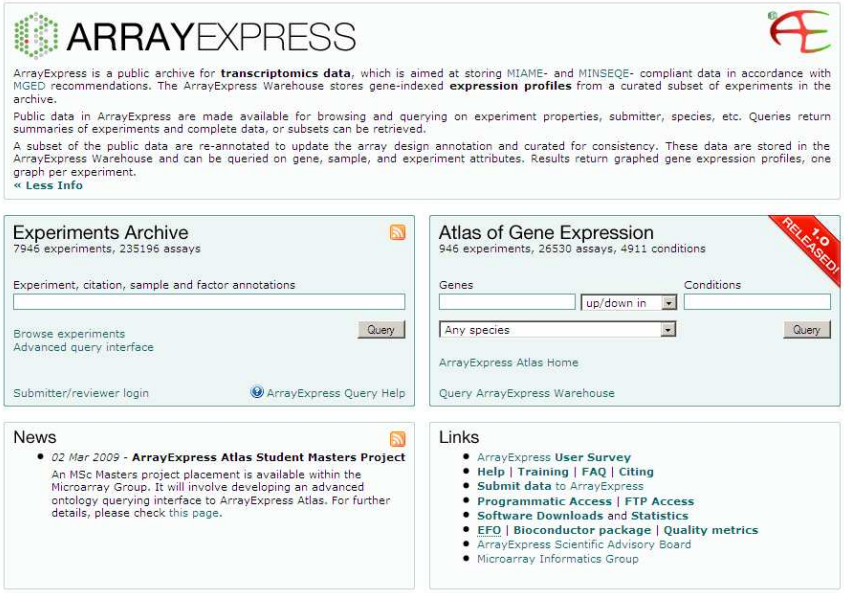



**Figura 2.5 Página inicial do Arrayexpress [http://www.ebi.ac.uk/microarray-as/ae/].**

## 2.2.2. NCBI

O NCBI (*Centro Nacional para a Informação de Biotecnologia*) (Figura 2.6) é parte do NML (*Biblioteca Nacional de Medicina*), uma filial dos institutos nacionais da saúde dos Estados Unidos. Fica situado em Bethesda, Maryland e foi fundado em 1988 [36].

O NCBI contém mais de 30 bases de dados. No entanto as que se tornam relevantes para a concepção deste trabalho são: *PubMed* [10], *Entrez Gene* [11] e *OMIM* [37].


**Figura 2.6 Página com informação das bases de dados presentes no NCBI [http://www.ncbi.nlm.nih.gov/].**

O Entrez Gene é uma base de dados de genes (Figura 2.7). O identificador de gene do Entrez Gene é um número inteiro único por gene. Todos os identificadores são integrados com o sistema de *queries*, *LinkOuts* e acesso por utilitários do *Entrez*, permitindo assim um acesso externo rápido e flexível [36].

A informação mantida nesta base de dados inclui nomenclatura, localização no cromossoma, produtos de genes e seus atributos (por exemplo, interacções proteicas), marcadores associados, fenótipos, sequências, e ainda uma grande quantidade de *links* para as citações, detalhes sobre variações das sequências, relatórios de expressão, homólogos, conteúdos e domínio de proteínas e acesso para bases de dados externas.

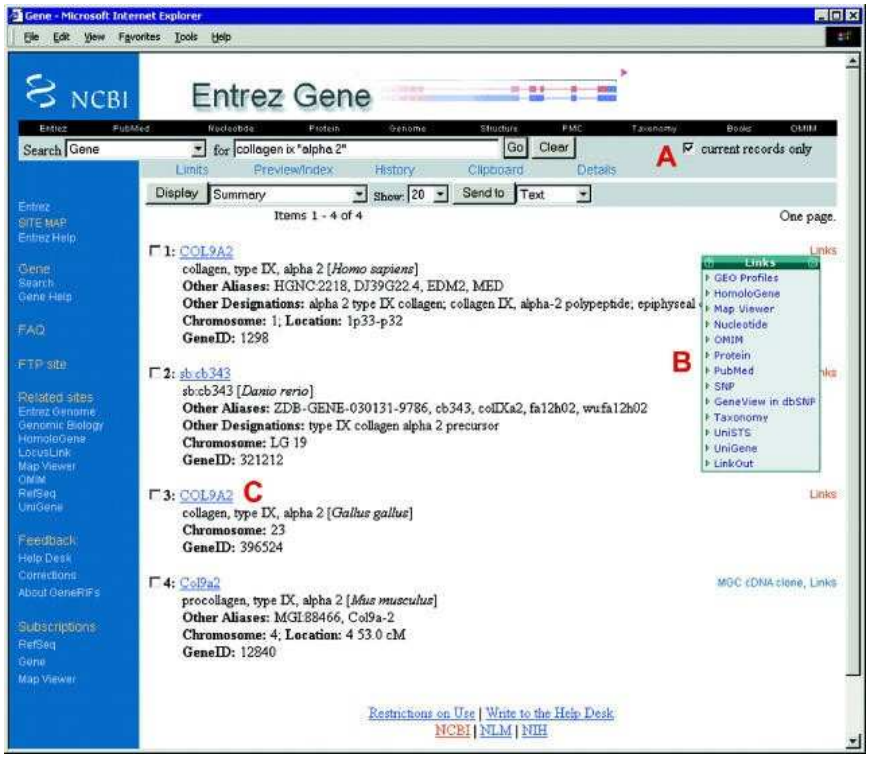



**Figura 2.7 Resumo de uma pesquisa do Entrez Gene [36].**

Os dados do *Entrez Gene*, são uma mistura de dados manualmente validados e análises computacional.

O *Entrez Gene* é uma parte fundamental da representação das informações específicas dos genes no NCBI. A informação sobre os genes tem interligação com outros recursos do NCBI [36], tais como o *BLAST*, *Geo*, *HomoloGene*, *Map Viewer*, *UniGene* e *UniSTS*. Por exemplo, os nomes associados a *GeneIDs* são utilizados no *HomoloGene*, *Map Viewer*, *UniGene* e os *Mammalian Gene Collection* [11]. Existe ainda uma equipa que verifica os desajustamentos na representação dos genes, tornando-se, por isto, um dos recursos mais importantes e válidos sobre genes, para a comunidade científica.

O projecto de desenvolvimento do OMIM (*Online Mendelian Inheritance in Man*) foi iniciado no início de 1960, como um catálogo das desordens Mendelianas, intitulado por MIM (*Mendelian Inheritance in man*). A versão online, OMIM, foi criada em 1985 e dois anos depois ficou disponível na Internet e em 1995 o OMIM foi desenvolvido para a *World Wide Web*.

O OMIM é usado primeiramente pelos médicos, mais propriamente, por investigadores na área da genética, e por estudantes na área avançadas de ciências e medicina (Figura 2.8) [37].

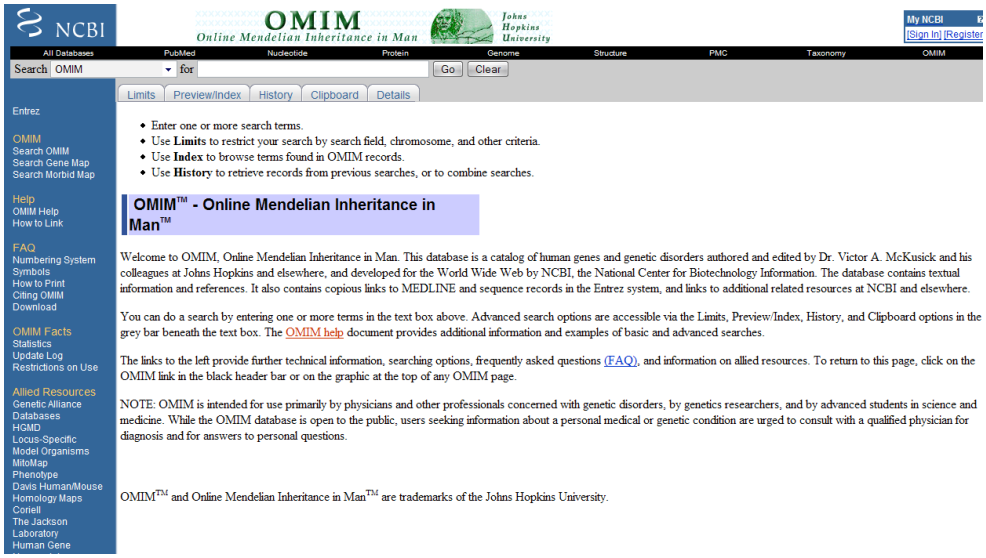



**Figura 2.8 Página inicial do Omim [http://www.ncbi.nlm.nih.gov/omim/].**

O OMIM é uma base de dados abrangente que contém os genes, os genótipos e fenótipos do humano. Os textos referenciados no OMIM contêm informações sobre todas as doenças mendelianas conhecidas e sobre mais de 12.000 genes. O OMIM incide sobre as relações entre genótipos e fenótipos. A base de dados do OMIM é actualizado diariamente e as entradas contêm *links* para outras bases de dados.

O *PubMed* é um serviço que inclui mais de 17 milhões de citações da MEDLINE e outros jornais de ciência e artigos biomédicos, que datem desde de 1950s. Tem ainda *links* para artigos completos e outros recursos relacionados [10]. Podemos ver a visualização de um artigo no *PubMed* na Figura 2.9.

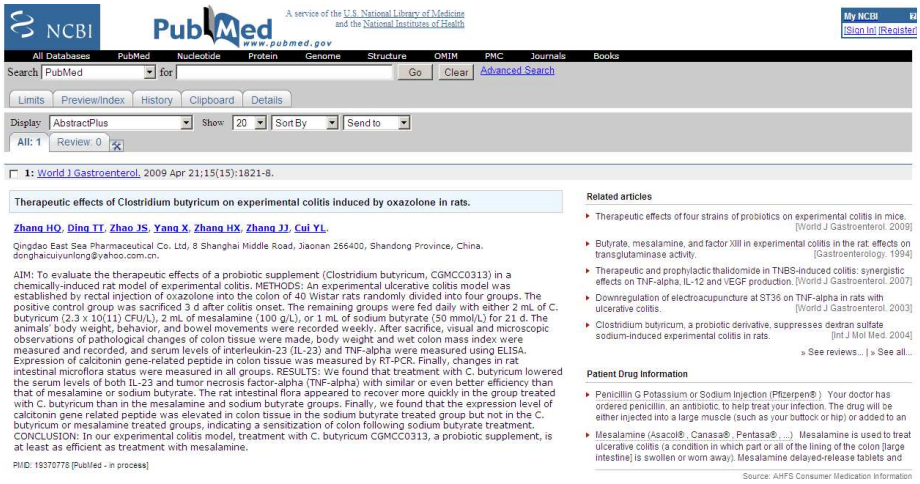



**Figura 2.9 Visualização de um artigo no Pubmed [http://www.ncbi.nlm.nih.gov/pubmed/].**

### 2.2.3. *KEGG*

A Base de Dados KEGG (*Kyoto Encyclopedia of Genes and Genomes*) foi iniciada em 1995, pelo programa de genoma humano Japonês. De acordo com os construtores, o *KEGG* é "Representação computorizada" de um sistema biológico (Figura 2.10) [38]. A Base Dados pode ser utilizada para modelação e simulação, procura e retorno de dados.

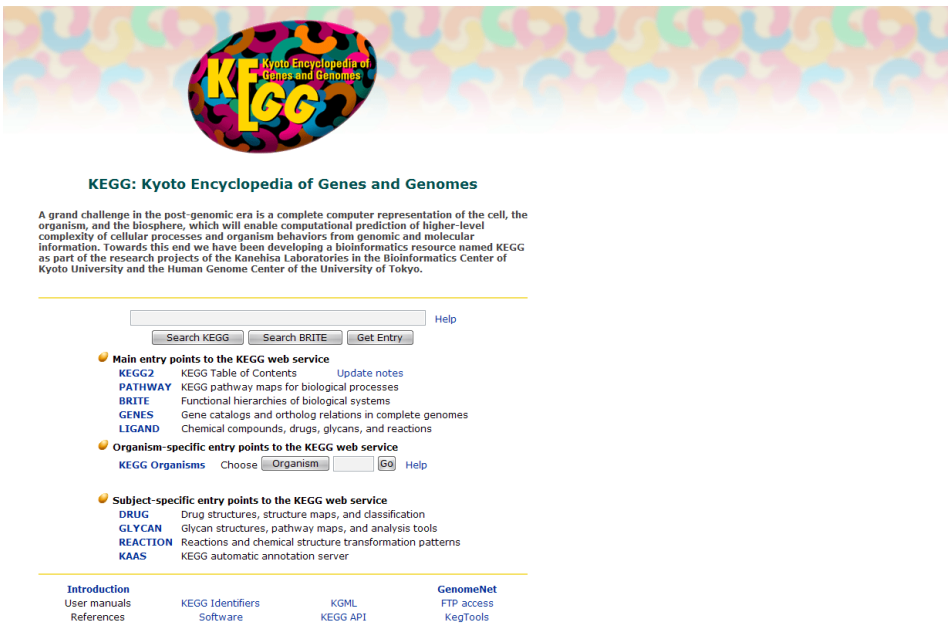



**Figura 2.10 Página de apresentação do *KEGG* [http://www.genome.jp/kegg/].**

O *KEGG* é constituído por 4 principais bases de dados, como podemos ver na Figura 2.11: *KEGG Pathway, KEGG Genes, KEGG Ligand* e *KEGG Brite* [39].

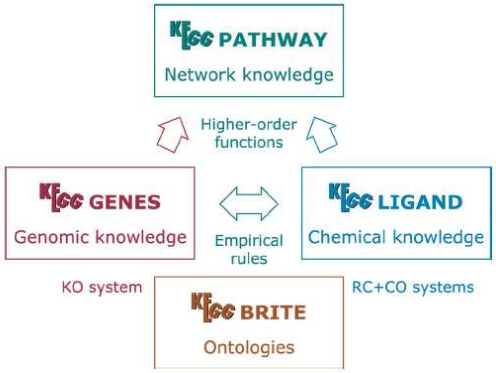



**Figura 2.11 Arquitectura interna do *KEGG* [39].**

O *KEGG* contém informação sobre redes de interacções moleculares, tais como vias metabólicas (Figura 2.12) (esta é a base de dados do *Pathway*), a informação sobre os genes e as proteínas gerados por projectos do genoma (inclui base de dados do gene) e a informação sobre compostos bioquímicos e reacções (inclui bases de dados de composto e de reacção).

A base de dados do *KEGG Pathway* é uma colecção de vias metabólicas desenhadas manualmente, contendo o processamento da informação genética, processamento da informação ambiental, como a transdução de sinais, bem como outros processamentos celulares e doenças humanas. Durante os últimos 2 anos, tem aumentado significativamente o número de vias metabólicas para regulação, tradução de sinais, ligação e interacção de receptores, comunicação das células, todos baseados em extensa pesquisa da literatura publicada [9].

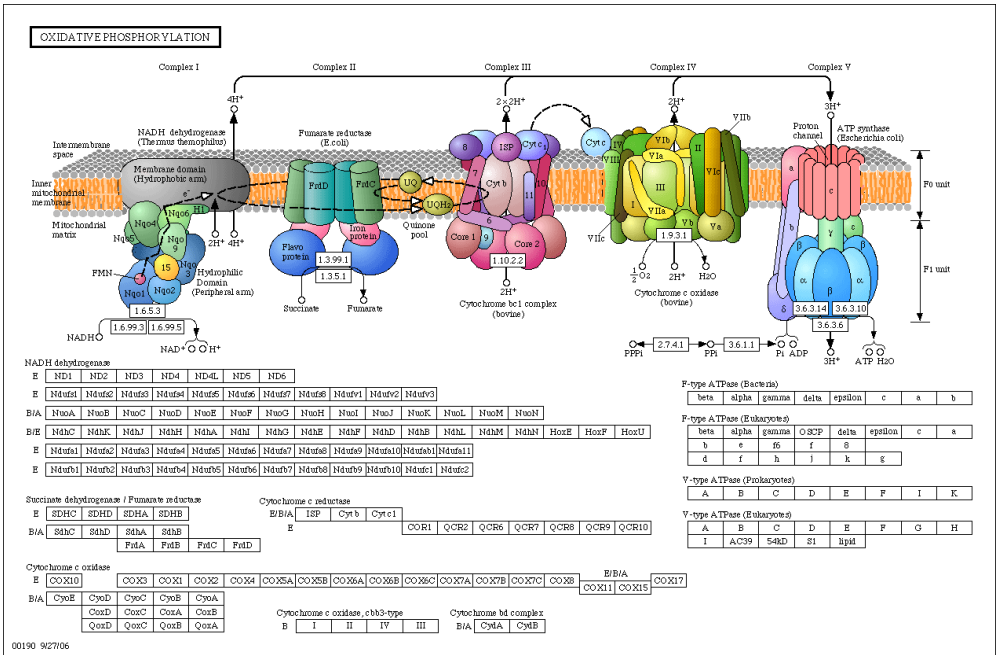



**Figura 2.12 Representação de uma via metabólica do *KEGG* [http://www.genome.jp/kegg/].**

O *KEGG Brite* (Figura 2.13) é uma colecção binária e hierárquica com dois objectos inter-relacionados correspondentes aos dois tipos de gráficos: para automatizar as interpretações funcionais 30 associações com o *KEGG Pathway* e para ajudar na descoberta tem regras empíricas envolvendo interacções ambientais do genoma. Actualmente, o *KEGG* está focado na estruturação hierárquica do conhecimento sobre aspectos funcionais e químicos do genoma [40].

O *KEGG Brite* inclui ortólogos/parálogos de grupos de genes, a classificação de reacção (RC) sistema de reacções bioquímicas e outras classificações de compostos químicos e drogas. O *KEGG*

tem planeado alargar o conhecimento dos KO para incluir a definição de módulos funcionais nas vias metabólicas existentes e desenvolver ontologias [40].


**Figura 2.13 Árvore de ortologias do *KEGG* [http://www.genome.jp/kegg/].**

## 2.2.4. *Gene ontology* (*GO*)

O projecto *Gene Ontology* (*GO*) é um esforço cooperativo que vem responder às necessidades de descrever de forma consistente os produtos dos genes nas diferentes bases de dados (Figura 2.14). Os vocabulários controlados (ontologias) que descrevem os produtos dos genes independentemente da espécie, estão divididos em três classes: processos biológicos, componentes celulares e funções moleculares. Há três aspectos separados neste esforço: primeiramente, escrever e manter as ontologias; em segundo, fazer ligações transversais entre as ontologias, os genes e os produtos dos genes nas bases de dados que cooperam com o projecto (GO) e, em terceiro lugar, desenvolver as ferramentas que facilitam a criação, manutenção e uso das ontologias [8].

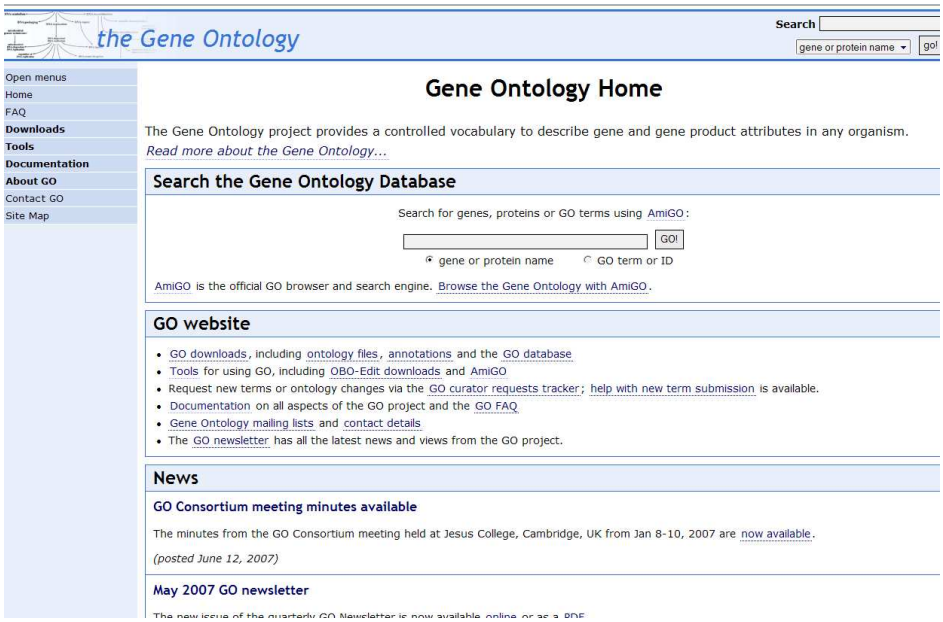



**Figura 2.14 Página inicial do *Gene Ontology* [http://www.geneontology.org/].**

Os vocabulários controlados são estruturados de modo a que o utilizador possa navegar pelos diferentes níveis: por exemplo, pode-se usar o GO para encontrar todos os produtos do gene no genoma de um rato que são usados na transmissão de um sinal (Figura 2.15) [8].

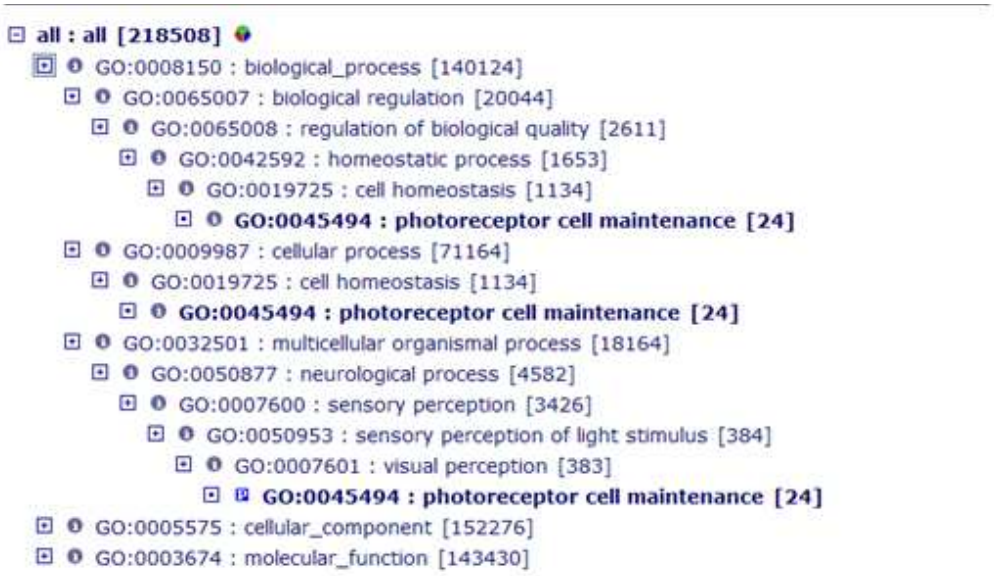



**Figura 2.15 Vista em árvore da ontologia [http://www.geneontology.org/].**

## 2.2.5. Outros

Para além das fontes de dados que foram referidas, torna-se relevante falar um pouco mais sobre as fontes de dados de homologias como o *ProDom* [28], *TIGRFAMS* [41] , *Prosite* [29] e *Prints* [27].

O *ProDom* é uma abrangente base de dados de famílias de domínios de proteínas geradas a partir da comparação global de todas as sequências de proteínas (Figura 2.16). Contém a visualização tridimensional de domínios e estruturas. Além disso, tem desenvolvido *ProDom-SG*, o *ProDom* que se dedica a selecção dos candidatos para proteínas estruturais e genómicas.

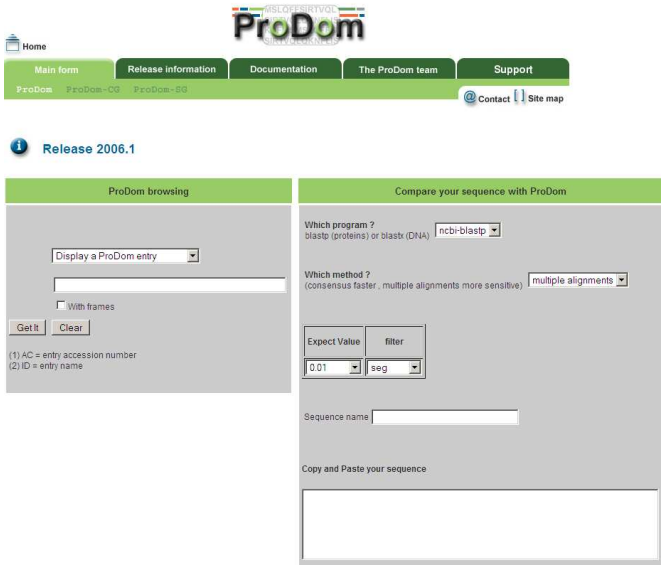



**Figura 2.16 Página inicial do ProDom [http://prodom.prabi.fr/prodom/current/html/home.php].**

O TIGRFAMs é uma colecção de famílias de sequências múltiplas de proteínas manualmente validadas. Contém informação destinada a apoiar a identificação automática e funcionamento das proteínas por sequência de homologia (Figura 2.17), baseado em HMMs (*Hidden Markov Models*). A classificação de família equivalente é sempre que possível, complementada pela classificação de ortólogos, superfamílias, domínio ou motifs. Ele fornece a informação mais adequada para atribuição automática de funções específicas às proteínas e projectos de sequenciação de genomas em grande escala.

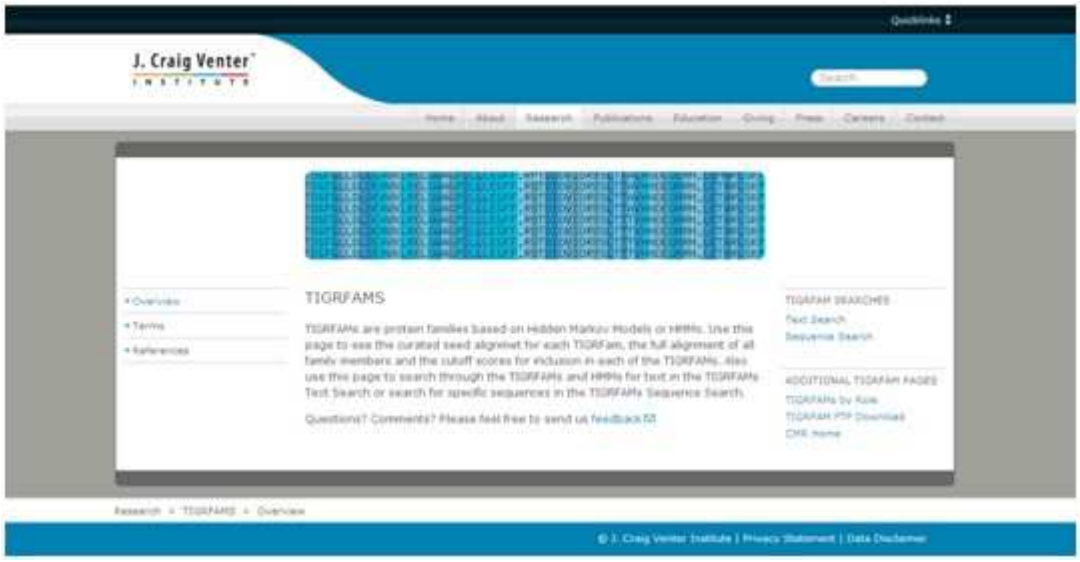



**Figura 2.17 Página inicial do Tigrfams [http://www.jcvi.org/cms/research/projects/tigrfams/overview/].**

O *Prosite* é uma base de dados de famílias e domínios de proteínas (Figura 2.18). É baseado no conceito de que embora existam um grande número de proteínas diferentes, a maioria delas podem ser agrupadas, num número limitado de famílias, com base nas similaridades que existem na sua sequência. Proteínas ou domínios de proteínas pertencentes a uma determinada família geralmente partilham atributos funcionais e derivam de um mesmo ancestral.

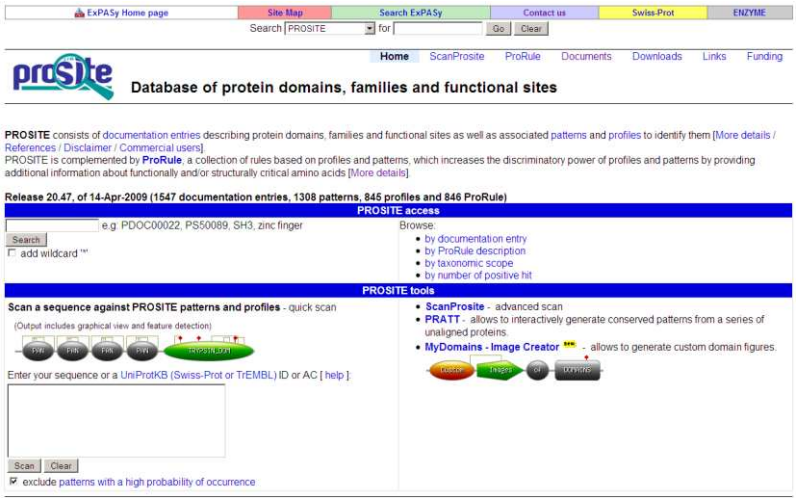



**Figura 2.18 Página inicial do ProSite [http://www.expasy.ch/prosite/].**

O *Prosite* é complementado pelo *ProRule*, que é um conjunto de regras com base em perfis e padrões, o que aumenta o poder discriminatório dos perfis e padrões, fornecendo informações adicionais sobre a funcionalidade e/ou aminoácidos estruturalmente críticos.

O *Prosite* contém actualmente padrões e perfis específicos para mais de um milhar de famílias ou domínios de proteínas. Cada uma destas assinaturas fornece documentação com informações básicas sobre a estrutura e função destas proteínas.

A base de dados do *PRINTS* é um resumo das impressões digitais dos *motifs* das proteínas (Figura 2.19). Cada *PRINT* é definida usando processos de pesquisa e análise de sequência como o ADSP ou VISTAS. Existem dois tipos de *PRINT*, simples ou complexos: os simples são geralmente um *motif*, os complexos são um conjunto de *motifs*.

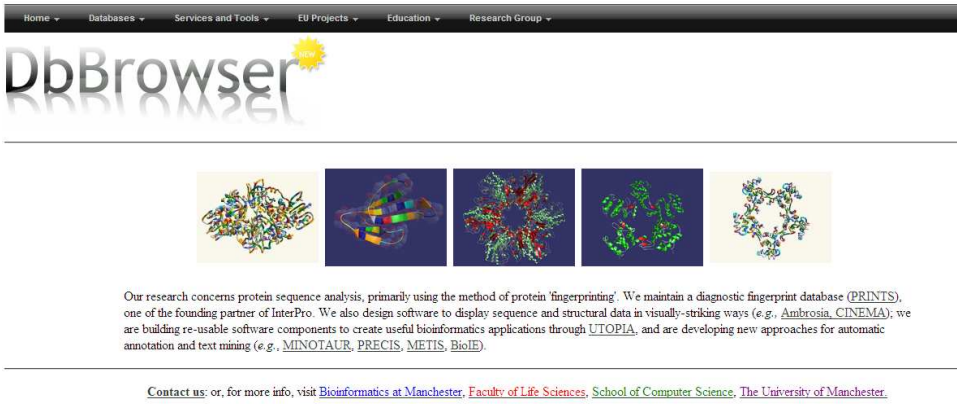



**Figura 2.19 Página inicial do Prints [http://www.bioinf.manchester.ac.uk/dbbrowser/index.php].**

## 2.2.6. Data *warehouses* em biologia molecular

Tal como referido anteriormente na última versão do [*Nucleic Acids Research]*, existem mais de 1000 bases de dados no domínio da biologia molecular [42].

Algumas das fontes dados mais relevantes da biologia molecular são as descritas anteriormente. No entanto, a integração de dados de várias destas fontes é importante por dois motivos:

- Porque os dados sobre uma entidade biológica podem estar dispersos por várias bases de dados, como por exemplo, de um gene, em que a sequência de nucleótidos é armazenada no *GenBank* [43], as vias metabólicas no *KEGG Pathway* [44] e os dados de expressão génica no *ArrayExpress* [34]. Obter uma visão unificada dos dados é fundamental para compreender o papel do gene.

- Segundo, consiste no facto de diferentes bases de dados conterem informações redundantes ou sobreposição de informação [45]. Isto pode ser detectado por comparação directa das bases de dados.

A maior parte dos dados armazenados nas referidas bases de dados está publicamente disponível na *Web* com interfaces personalizados, ou com texto e arquivos *XML* [46]. Para a obtenção desses dados tem de se aceder a cada base de dados independentemente, fazer *download* e analisar os processos de extracção de informação, finalmente juntar todos os resultados de um conjunto de dados e unificá-los de forma coerente.

Nos últimos anos, têm sido efectuados vários esforços para simplificar o processo de integração de dados de múltiplas fontes. Um desses, *BioWarehouse* [47], contém dados de múltiplas fontes, incluindo vias metabólicas e enzimas. O *BioWarehouse* utiliza um esquema de dados orientado a tipos de dados predefinidos, o que significa que a adição de novos tipos de dados implica adicionar novas tabelas e os métodos de consulta. Esta base de dados foi projectada tendo uma orientação mais para espécies procariotas do que para eucariotas.

Uma visão diferente foi aplicada no *BioCoRE* [48] que utiliza uma abordagem mais flexível para integrar os dados. Segundo os autores, o sistema permite o armazenamento de quase todos os processos bioquímicos. Uma desvantagem é a alta complexidade do modelo proposto que contém mais de 200 classes.

Uma terceira abordagem tem sido aplicada pelo *Biozon* que suporta os dados baseados num meta-modelo hierárquico [49]. Uma vez que o esquema é geral, em cada relação do *Biozon* o meta-modelo é explicitamente armazenado na base de dados. Como consequência, o actual exemplo contém cerca de 6,5 mil milhões de relações, que diminuem o desempenho. O *Biozon* está disponível publicamente através de um intuitivo e fácil interface *Web*, mas não é possível fazer o *download* da base de dados, a fim de instalar localmente. Temos ainda o *Bio2RDF*, uma data *Warehouse* que faz a conversão de conhecimento bioinformática em RDF, embora tenha informação limitada de entidades biológicas, como vias metabólicas, homologias e enzimas. Apenas contém dois organismos, humano e rato, o que vem limitar a capacidade de utilização da base de dados.

Os sistemas de dados anteriores, apresentam diferentes abordagens para resolver o mesmo problema, integrar dados de diferentes fontes. As limitações encontradas reflectem a dificuldade de obter um esquema simples, mas abrangente, que é capaz de acomodar a heterogeneidade do

domínio biológico e manter um bom nível desempenho. Neste âmbito, está a ser desenvolvido pelo grupo de bioinformática da Universidade de Aveiro a base de dados GeNS, que se obtém e unifica as fontes de dados mais relevantes no contexto do presente trabalho.

#### *2.2.6.1. GeNS*

O esquema de base de dados do GeNS [*Tese de MSc João Pereira*], tira partido da escalabilidade e flexibilidade para armazenamento de dados biológicos. Para armazenar os dados fisicamente o GeNS usa um modelo geral que certifica a escalabilidade e a flexibilidade da base de dados. Para atingir o modelo físico é representado num meta-modelo em que todas as entidades e relações são especificadas.

A Figura 2.20 contém o meta-modelo, onde o gene desempenha um papel central. Relacionados com cada gene existe uma rede de tipos de dados que relaciona os termos com as bases de dados às quais estão associados (Figura 2.21). A adição de novas bases de dados implica a adição de novos tipos de dados e requer apenas mudanças no meta-modelo e não no modelo físico.

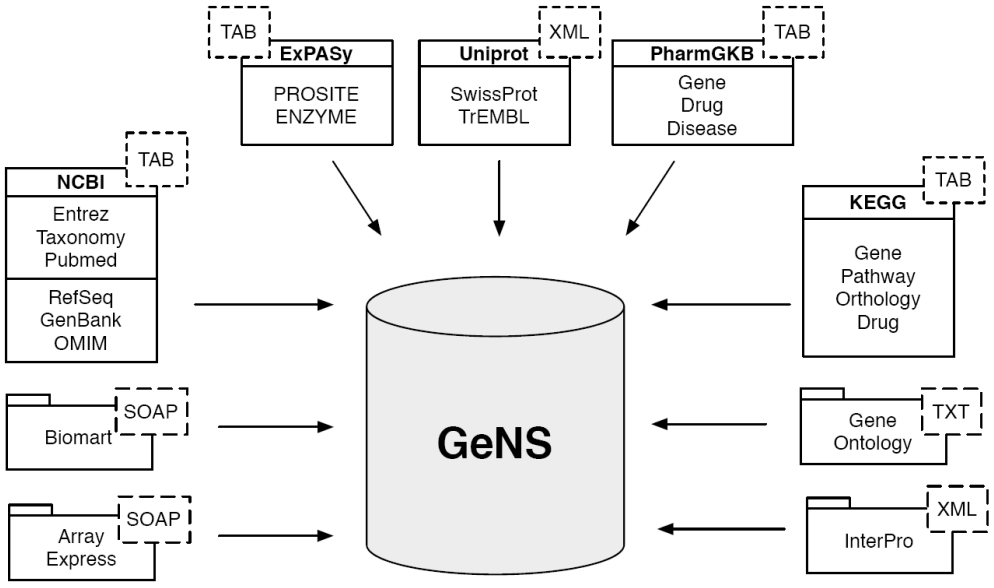



**Figura 2.20 Esquema da representação das bases de dados integradas no GeNS.**

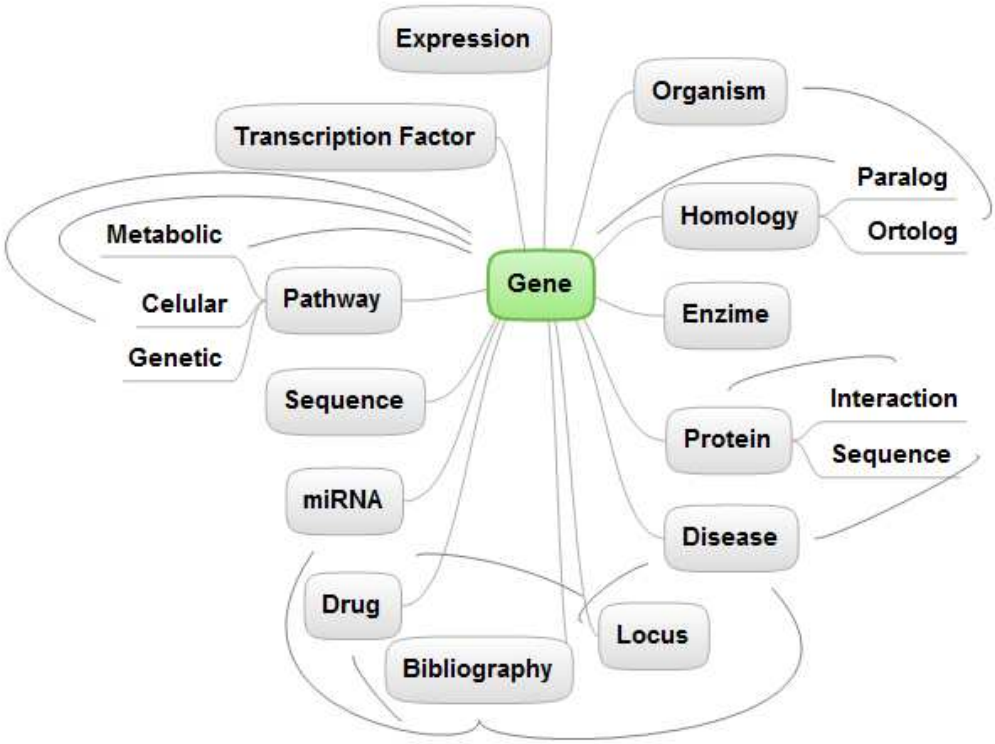



**Figura 2.21 Modelo de relações presentes no GeNS.**

### Modelo físico da Base de dados

A Figura 2.22 representa o modelo físico de dados. Porque precisávamos de armazenar explicitamente todas as relações entre genes e proteínas, devido a fins aplicativos foi mudado o papel central do gene no meta-modelo para a proteína no modelo físico.

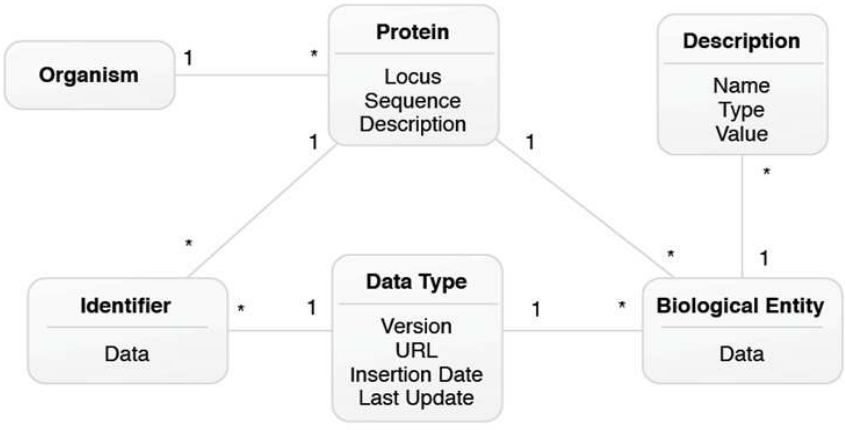



**Figura 2.22 Modelo físico do GeNS.**

Na sequência, iremos descrever em detalhe os conceitos aplicados na elaboração desta base de dados.

***Organism***: Guarda informação taxonómica; cada entrada corresponde a um organismo, e para estes contém o nome a abreviatura e o identificador taxonómico. E esta tabela constitui a raíz do modelo hierárquico.

***Protein***: Esta tabela armazena informações da relação entre organismo e proteína, bem como, informação que inclui a localização do gene no cromossoma do gene associado à proteína e a sua sequência.

***Identifier***: Contém todos os sinónimos de genes e proteínas, nomes alternativos e identificadores para cada entrada.

***Datatype***: Contém uma lista com todos os tipos de dados extraídos de bases de dados externas, abrangendo ambos os identificadores e entidades biológicas.

***BioEntity***: Esta tabela armazena identificadores únicos pertencentes às entidades biológicas associadas a uma determinada proteína, o que inclui, entre outras coisas, identificadores via de sinalização, dados de expressão génica, ontologia, etc. Dados que descrevem estas entidades biológicas, serão colocados na tabela *Description*.

Esta organização hierárquica não só simplifica o esquema de dados, ou seja, torná-la mais fácil de compreender e manter, mas também melhora o desempenho do sistema, simplificando o acesso aos dados. O sistema também é muito flexível devido à forma de mapeamento de dados para proteínas: cada entidade biológica na tabela *BioEntity* é única e, graças a uma associação na tabela *ProteinBioEntity*, cada proteína contém múltiplas conexões com cada entidade biológica, assegurando assim a escalabilidade e o desempenho do sistema, juntamente com todos os benefícios proporcionados pelo modelo hierárquico.

**Caso de utilização**

O exemplo da Figura 2.23 demonstra um dos muitos cenários possíveis presentes no GENS: um investigador pretende obter a rede de conceitos relacionados com um gene “sce:Q0085”. O sistema começa por determinar o identificador interno de proteína através da tabela *Identifier*. Com este identificador, podemos agora determinar a nomes alternativos do gene e proteína.

Posteriormente, o sistema irá verificar o correspondente organismo; neste caso particular, já sabemos a resposta devida às três primeiras letras do identificador (*sce*, o acrónimo de *Saccharomyces cerevisiae*), mas este facto não vai afecta o processo. Para isso o GeNS olha para a

tabela *Protein*, e retira o identificador taxonómico presente na tabela, para identificar o organismo vai a tabela *Organism*. Na tabela a *Protein* também é possível encontrar a localização do gene no cromossoma e a sua sequência.

Após este procedimento, é possível mapear no GeNS cada entidade biológica associada à nossa pré-determinada proteína, observando-se a tabela *ProteinBioEntity* (que contém todas as relações entre os dois). Isto permite ao GeNS recuperar as entidades biológicas na tabela *BioEntity*, que por sua vez, contém homologia, bibliografia, dados de expressão, de ontologia, vias de sinalização, entre outros.

Finalmente, pode ser obtido mais detalhe sobre cada entidade biológica através da observação de sua descrição na tabela *BioEntityDescription*.

Para alargar este exemplo, o investigador pretende obter todos os outros genes relacionados com a via metabólica do *KEGG* "*sce00190*", onde o gene "*sce:Q0085*" estava inicialmente presente. Para isso ele procura todas as proteínas que contenham uma relação com a tabela *BioEntity* que corresponda a via metabólica pretendida.

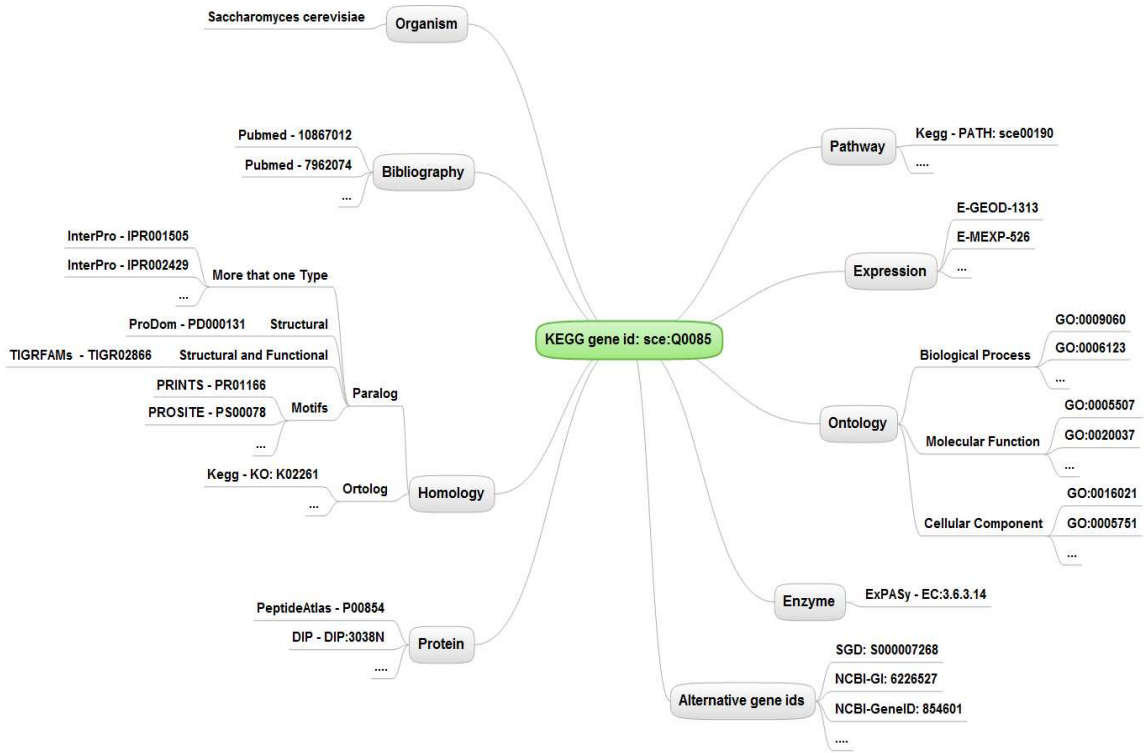



**Figura 2.23 Exemplo de utilização do GeNS com o gene *"sce:Q0085"*.**

## 2.3. Aplicações *Web* para Interpretação de Dados Biológicos

Existe um grande número de aplicações *Web* que integram dados biológicos e entre estas existe um subconjunto de aplicações que se focam também na interpretação de dados biológicos. Este subconjunto de aplicações tem vindo a crescer e a tornar-se mais relevante, devido ao facto de auxiliar o trabalho dos biólogos, na obtenção de respostas a um dado problema.

Ferramentas como o *Onto-Express* [2], *FatiGO* [3, 50], *GOTree* [51], *GoMiner* [4], *PathwayExplorer* [6], *CARGO* [52], *DiseaseCard* [53] e a versão anterior do *GeneBrowser* [7], apresentam uma perspectiva semelhante ao que se pretende com a nova versão do *GeneBrowser*.

O *Onto-Express* [2] é capaz de construir perfis funcionais usando o *Gene Ontology* (GO), calculando a significância estatística de cada categoria. Uma outra ferramenta muito semelhante é o *GOTree* [51] e o *GoMiner*, que faz também cálculo da significância estatística das classes do GO. O *Babelomics* [54-56] contém um conjunto de ferramentas para análise funcional de genes, entre as quais se encontra o *FatiGO*. Estas ferramentas incluem análise ontológica (GO), homologica, vias de sinalização e cromossómica.

Ferramentas como a versão anterior do *GeneBrowser*, *PathwayExplorer*, *DinamicFlow*, *CARGO* e o *DiseaseCard* utilizam uma abordagem distinta das anteriores. Apenas procuram mostrar informação ao utilizador, estabelecendo relações e métodos de visualização de informação, sem calcular qualquer significância estatística.

O conjunto de aplicações que apresenta como solução uma análise funcional utilizando métodos estatísticos apresenta vantagens, pois vem diminuir o número de classes funcionais que o utilizador tem de analisar para obter uma resposta. No entanto, ferramentas como o *FatiGO*, *GOTree*, ou *Onto-Express* são confusas e lentas na análise dos dados, devido a não apresentarem um ambiente integrado que mostre num acesso único acesso toda a informação ao utilizador.

O que se pretende com a nova versão do *GeneBrowser* é aproveitar as vantagens de uma análise funcional utilizando métodos estatísticos, sem se descuidar no desempenho e simplicidade, tornando o ambiente fácil de usar e rápido no processamento de informação.

## 2.4. Sumário

O mapeamento de conceitos e relações na biologia molecular exige tipicamente a agregação de inúmeras fontes de dados. Dado o aspecto singular de cada fonte de dados, são requeridas diferentes estratégias computacionais para acesso e extracção de dados. A integração de dados biológicos já existe a algum tempo embora não exista uma solução conceptual. Existem várias aplicações com esta finalidade, no entanto cada uma usa a sua abordagem, de modo a responder a uma ou várias perguntas biológicas.

# Capítulo 3 - Arquitectura de um Sistema para Interpretação de Dados Biológicos

Neste capítulo pretende-se definir a arquitectura típica de um sistema, baseado em conceitos de programação *Web 2.0*, que suporte a integração de dados biológicos de diferentes fontes e que permita ao utilizador a interpretação dos resultados obtidos experimentalmente.

O objectivo é permitir que o utilizador insira os seus *Datasets*, tipicamente um conjunto de identificadores de genes, e obtenha de uma forma simples e rápida os resultados. Todos os resultados deverão conter informação sobre ontologia, homologia, vias de sinalização, bibliografia, informação descritiva de cada gene, informação de expressão génica e informação sobre a localização dos genes no cromossoma.

Uma particularidade importante desta aplicação é para além da integração de dados biológicos, também a inclusão de modelos estatísticos, que vem dar relevância qualitativa há integração de dados.

O sistema foi construído de forma a ser escalável e flexível, a nível de componentes que utiliza, como a nível de serviços e fontes de dados a serem integradas. A complexidade da arquitectura é ainda aumentada devido ao facto de se tratar de uma aplicação *Web*: não seria instalada num computador próprio para uso particular e tinha de estar sempre online, disponível a qualquer momento e a partir de qualquer local e funcionar nos *browsers* mais utilizados.

## 3.1. Web 2.0

O termo *Web 2.0* foi popularizado pela "*O'Reilly Media*" e pela "*MediaLive International*", como conceito emergente numa série de conferências que tiveram início em Outubro de 2004 (*O'Reilly, 2005*).

O conceito *Web 2.0* é a denominação de uma segunda geração de serviços *online* caracterizada por potencializar de forma alargada os mecanismos de publicação, compartilhando e organizando as informações, além de ampliar os espaços para a interacção entre os participantes do processo.

O *Web 2.0* refere-se não apenas a uma combinação de técnicas informáticas (*Web Services*, linguagem *Ajax*, *Web syndication*, etc.), mas também a um determinado período tecnológico, a um conjunto de novas estratégias e a processos de comunicação mediados pelo computador [57].


**Figura 3.1 Aplicações Web 2.0.**

## 3.1.1. AJAX

AJAX (*Asynchronous Javascript And XML*) é um conjunto de metodologias que tornam as páginas mais interactivas com o utilizador, recorrendo a pedidos assíncronos de informações. O AJAX é também uma iniciativa para a construção de aplicações *Web* dinâmicas e criativas. AJAX é um conjunto de tecnologias conhecidas trabalhando juntas, cada uma fazendo a sua parte, e oferecendo novas funcionalidades [58].

O AJAX incorpora no seu modelo várias características:

- Apresentação baseada em padrões, usando XHTML e CSS;
- Exposição e interação dinâmica usando o DOM;
- Intercâmbio e manipulação de dados usando XML e XSLT;

- Recuperação assíncrona de dados usando o objeto *XMLHttpRequest*;
- *JavaScript*.

### 3.1.2. Javascript

JavaScript é uma linguagem de programação criada pela *Netscape* em 1995 para responder as necessidades de validação de formulários no lado cliente e interacção com a página tornando as páginas HTML dinâmicas DHTML (*Dynamic HTML*). O DHTML nasceu da união do *JavaScript* com o CSS (*Cascading Style Sheets*) e com este foi possível modificar dinamicamente os estilos dos elementos das páginas HTML [59]. Devido a esse facto é que o *JavaScript* foi construído como uma linguagem de *script* [60, 61]. Podemos ver um exemplo de *JavaScript* na Figura 3.2.

```
<script>
// We can't manipulate the DOM until the document
// is fully loaded
window.onload = function(){

    // Find all the <li> elements in the document
    var li = document.getElementsByTagName("li");

    // and add a ared border around all of them
    for ( var j = 0; j < li.length; j++ ) {
        li[j].style.border = "1px solid #000";
    }

    // Locate the element with an ID of 'everywhere'
    var every = document.getElementById( "everywhere" );

    // and remove it from the document
    every.parentNode.removeChild( every );

};
</script>
```

**Figura 3.2 Exemplo de Javascript que altera o estilo da border de todos os elementos <li> list item e remove os filhos do elemento que tem id "everywhere".**

### 3.1.3. XML

O XML (*eXtensible Markup Language*) é um subtipo do SGML (*Standard Generalized Markup Language*) recomendado pelo W3C (*The World Wide Web Consortium*) para gerar linguagens de marcas, e estruturas de dados. O propósito do seu desenvolvimento era facilitar a partilha de informações na *Internet* [62, 63].

As suas principais características incluem a separação do conteúdo da formatação, simplicidade e legibilidade tanto para humanos como para computadores, possibilidade de criação de marcas sem limitação, criação de arquivos para validação de estrutura DTD (*Document Type Definition*), interligação de diferentes bases de dados e concentração na estrutura da informação, e não na sua aparência.

O XML é considerado um bom formato para a criação de documentos com dados organizados de forma hierárquica, como se vê frequentemente em documentos de texto formatados, imagens vectoriais ou bases de dados. Temos um exemplo de XML na Figura 3.3.

```
<book>
  <booktitle> The Selfish Gene </booktitle>
  <author id = "dawkins">
        <name>
            <firstname> Richard </firstname>
            <lastname> Dawkins </lastname>
        </name>
        <address>
            <city> Timbuktu </city>
            <zip> 99999 </zip>
        </address>
  </author>
</book>
```

**Figura 3.3 Exemplo de uma estrutura de dados em XML.**

### 3.1.4. CSS

CSS é uma linguagem de estilos utilizada para definir a apresentação de documentos escritos numa linguagem de marcas, como HTML ou XML. A principal vantagem na sua utilização é garantir a separação entre o formato e o conteúdo de um documento [64].

Apesar das sucessivas actualizações dos *browsers*, o suporte das várias versões do CSS (1, 2 e 3) não é uniforme o que complica o desenvolvimento de aplicações. Podemos ver um exemplo de CSS na Figura 3.4.

```
body {
    background-color:#CCCCCC;
    font-family:Arial, Helvetica, sans-serif;
    font-size:12px;
    margin-top:0;
    margin-right:0;
    margin-bottom:0;
    margin-left:0;
}

div {
    width:300px;
    height:300px;
    border-top:1px solid #000000;
    border-right:1px solid #000000;
    border-bottom:1px solid #000000;
    border-left:1px solid #000000;
}

h1 {
    font-size:medium;
}

p {
    line-height:1.5em;
}
```

**Figura 3.4 Exemplo de CSS, onde é difinido o estilo base de uma pagína html.**

### 3.1.5. JSON

JSON (*JavaScript Object Notation*) é uma formatação para troca de dados. É um formato de texto independente de linguagens, apesar de usar convenções que são familiares às linguagens C, C++, C#, *Java*, *JavaScript*, *Perl*, *Python* e muitas outras. Estas propriedades aliadas a um menor impacto da estrutura no volume final dos dados (como XML), fazem com que JSON seja um formato ideal de troca de dados.

A simplicidade do JSON tem resultado num uso difundido, especialmente como alternativa ao XML em AJAX. As principais vantagens do JSON sobre XML na troca de dados é o facto de ser muito mais fácil escrever um analisador JSON com a ajuda da função *eval* e da quantidade de informação a ser transmitida ao cliente ser menor que em XML. Podemos ver um exemplo da formatação de uma estrutura de dados em JSON na Figura 3.5.

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
```

**Figura 3.5 Exemplo de uma estrutura de dados em JSON.**

### 3.1.6. Web services

Os *Web services* são componentes que permitem às aplicações enviar e receber dados em formato XML. Cada aplicação pode ter a sua própria "linguagem" que é traduzida para uma linguagem universal, o formato XML [65].

Para as empresas, os *Web services* podem trazer agilidade para os processos e eficiência na comunicação entre cadeias de produção ou de logística. Todas as comunicações entre sistemas passam a ser dinâmicas e seguras, pois não há intervenção humana.

Essencialmente, o *Web Service* faz com que os recursos da aplicação do *software* estejam disponíveis sobre a rede de uma forma normalizada [66].

## 3.2. Requisitos do Sistema

Os requisitos do sistema foram definidos e são apresentados nesta secção como um conjunto de funcionalidades. As várias funcionalidades pretendidas para o sistema que se desenvolveu encontram-se organizadas em nove grupos distintos, descritos na Tabela 3.1. Para facilitar a interpretação das funcionalidades do sistema são apresentados também diagramas de casos de utilização. Não é definida grande parte dos actores porque são sistemas externos que já foram definidos no capítulo 2 na subsecção fontes de dados.

**Tabela 3.1 Grupo de funcionalidades**

| Grupos | Descrição |
|---|---|
| **Controle de acesso** | Engloba toda a informação relativa aos utilizadores do sistema, desde a criação de conta pelo utilizador não registado, à sua eliminação pelo administrador. |
| **Dataset** | O *dataset* contém toda a informação a armazenar no sistema. Esta informação incorpora duas listas de genes, um organismo e outra informação relevante para o dataset. |
| **Descrição geral dos genes** | Este grupo de funcionalidades disponibiliza informação que descreve os genes e os seus produtos. Essa informação pode ser descritiva ou obtida pelo acesso a fontes de dados externas. |
| **Descrição ontológica** | A descrição ontológica engloba um conjunto de funcionalidades que se baseiam em métodos de selecção e visualização de informação relativos às ontologias em que os genes de um dataset estão envolvidos. |
| **Homologias** | Homologias engloba um conjunto de funcionalidades que vão desde a selecção da fonte de dados ao acesso a sistemas externos. |
| **Vias metabólicas** | Este grupo engloba um conjunto de funcionalidades que se baseia em modos de selecção de informação e acesso a fontes de dados externas. |
| **Localização dos genes nos cromossomas** | Localização dos genes nos cromossomas engloba um conjunto de funcionalidades que se baseia na construção de métodos de acesso a fontes de dados externas. |
| **Dados de experiências anteriores** | Dados de experiências anteriores contém um conjunto de funcionalidades que permitem explorar informação e acesso a fontes de dados externas. |
| **Bibliografia** | Bibliografia contém métodos de visualização, enriquecimentos de informação, pesquisa e acesso a fontes externas que contém mais informação. |

### 3.2.1. Controlo de acesso

Este conjunto de funcionalidade (Tabela 3.2 e a Figura 3.6) demonstra que o sistema deve ser construído de forma a poder ser usado sem o utilizador se registar, ou seja, temos três tipos de utilizador: administrador do sistema, utilizador não registado e utilizador registado.

**Tabela 3.2 Funcionalidades do grupo controlo de acesso.**

| Grupo | Funcionalidade | Descrição |
|---|---|---|
| Conta | Login | Opção apenas presente para utilizador registado. |
| | Mudar Password | Opção apenas presente para utilizador registado. |
| | Remover Utilizador | Apenas o administrador do sistema pode remover um utilizador |
| | Registar | Todos os utilizadores podem efectuar registo, não dependendo este de uma validação do administrador. |

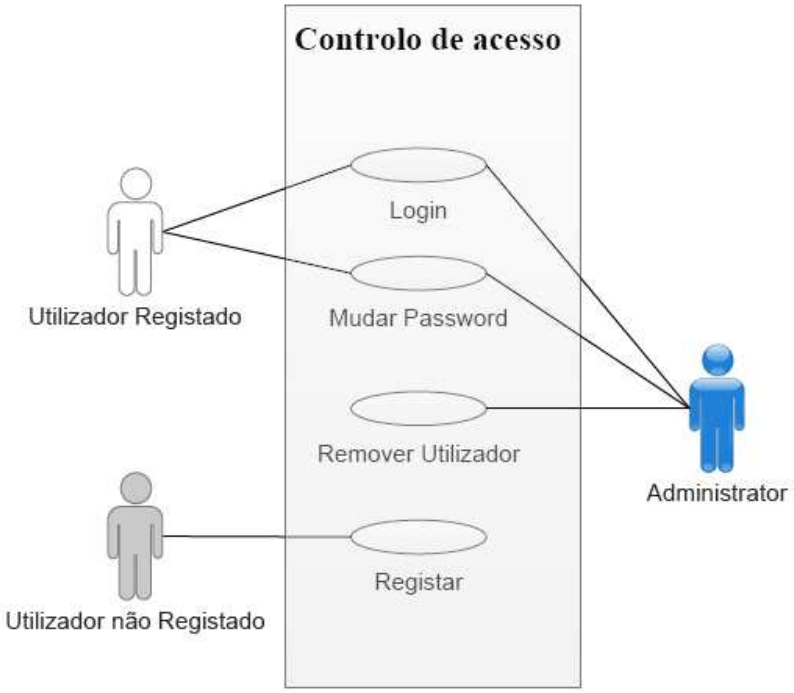



**Figura 3.6 Diagrama de casos de utilização do grupo controlo de acesso.**

## 3.2.2. Dataset

O grupo de funcionalidades da Tabela 3.3 e Figura 3.7, permite ao utilizador inserir um conjunto de dados (genes), mais propriamente, no processo de inserção, o utilizador tem de escolher um organismo, inserir uma primeira lista de genes, contendo os genes diferencialmente expressos e escolher comparar estes genes com todo o genoma do organismo ou inserir uma segunda lista com a qual pretende comparar a primeira lista de genes. Toda esta informação constitui a informação

presente num *dataset* e todos os identificadores de genes e organismos tem de ser validados pelo GeNS.

Um *dataset* pode ser editado, mas dentro deste apenas se pode editar a primeira lista de genes, genes diferencialmente expressos. Existem muitas outras operações que são realizadas sobre um *dataset*, apagar, gravar, mudar o nome e fazer *download* da informação contida neste.

Este grupo de funcionalidades distingue o utilizador registado do utilizador não registado, tendo o utilizador registado acesso a todas as funcionalidades, enquanto o utilizador não registado não tem acesso ao editar, apagar, guardar e *download* de informação relativa ao dataset.

Após inserção e processamento de um *dataset*, ou a selecção de um *dataset* por um utilizador registado, acedemos a todas as funcionalidades do *GeneBrowser* que inclui a descrição geral dos genes, descrição ontológica, vias metabólicas, homologias, localização do gene no cromossoma, dados de experiências anteriores e bibliografia. Em cada grupo de funcionalidades o utilizador acede aos dados do dataset seleccionado, e estes contêm diferentes métodos de selecção e exploração de dados.

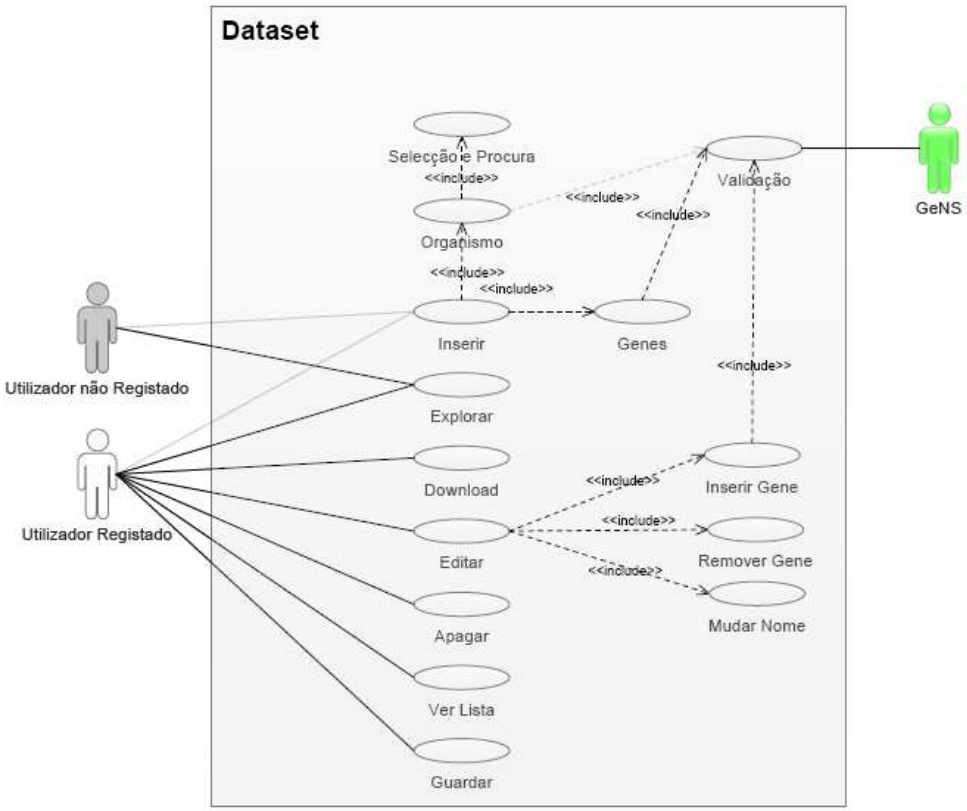



**Figura 3.7 Diagrama de casos de utilização do grupo dataset.**

**Tabela 3.3 Funcionalidades do grupo Dataset.**

| Grupo | Funcionalidade | Descrição |
|---|---|---|
| **Dataset** | Inserir | Todo o processo de criação de um novo *dataset*. |
| | Organismo | O organismo que o utilizador tem de inserir. |
| | Genes | As duas listas de genes a serem adicionadas pelo utilizador. |
| | Explorar | Consultar e utilizar a informação do dataset para outros estudos. |
| | Download | Fazer *download* da informação disponível num *dataset*. Opção apenas presente para utilizador registado. |
| | Gravar | Gravar um *dataset*. Opção apenas presente para utilizador registado. |
| | Apagar | Utilizador remove um d*ataset* previamente guardado. Opção apenas presente para utilizador registado. |
| | Listar Dataset | Consultar os datasets inseridos previamente pelo utilizador. Opção apenas presente para utilizador registado. |
| | Editar | Editar conteúdo de um *dataset*. Opção apenas presente para utilizador registado. |
| | Listar Genes | Ver a lista de genes de um *dataset*. Opção apenas presente para utilizador registados. |
| | Inserir Gene | Adicionar um gene a um *dataset* existente. Opção apenas presente para utilizador registado. |
| | Remover Gene | Remover um gene do *dataset*. Opção apenas presente para utilizador registado. |
| | Validar | Validação dos genes e organismos que o utilizador inserir. |

### 3.2.3. Descrição geral dos genes

Quanto à descrição geral dos genes (Tabela 3.4 e Figura 3.8) são apresentadas as funcionalidades que permitem ao utilizador consultar informação que descreve cada um dos genes e os seus produtos. Estão presentes também os métodos que permitem ao utilizador restringir a informação a vizualizar: *Select Display Info* permite a selecção dos campos de informação a serem disponibilizados de modo ao utilizador ter uma vista adaptada as necessidades, o *Filter Data*

permitem ao utilizador seleccionar os genes com base no cromossoma, ontologias, homologias e vias de sinalização.

**Tabela 3.4 Funcionalidades do grupo descrição geral dos genes.**

| **Grupo** | Funcionalidade | Descrição |
|---|---|---|
| **Descrição geral dos genes** | Select Display Info | Permite a selecção da informação a visualizar. |
| | Summary | Contém informação para sumário que descreve o gene. |
| | Homology | Contém informação sobre as classes homologicas. |
| | Gene Ontology | Contém informação as classes de ontologia |
| | Structure | Contém informação sobre a estrutura. |
| | Sequence | Contém informação sobre a sequência. |
| | Pathway | Contém informação sobre vias metabólicas. |
| | References | Contém informação para referências externas. |
| | Filter Data | Permite ao utilizador a filtragem de informação. |
| | Ontology Filter | Permite ao utilizador a filtragem de informação por classe de ontologia. |
| | Locus Filter | Permite ao utilizador a filtragem de informação por cromossoma. |
| | Pathway Filter | Permite ao utilizador a filtragem de informação por identificador de via metabólica. |
| | Homology Filter | Permite ao utilizador a filtragem de informação por classe de homologia. |
| | Export | Fazer *download* de informação. |

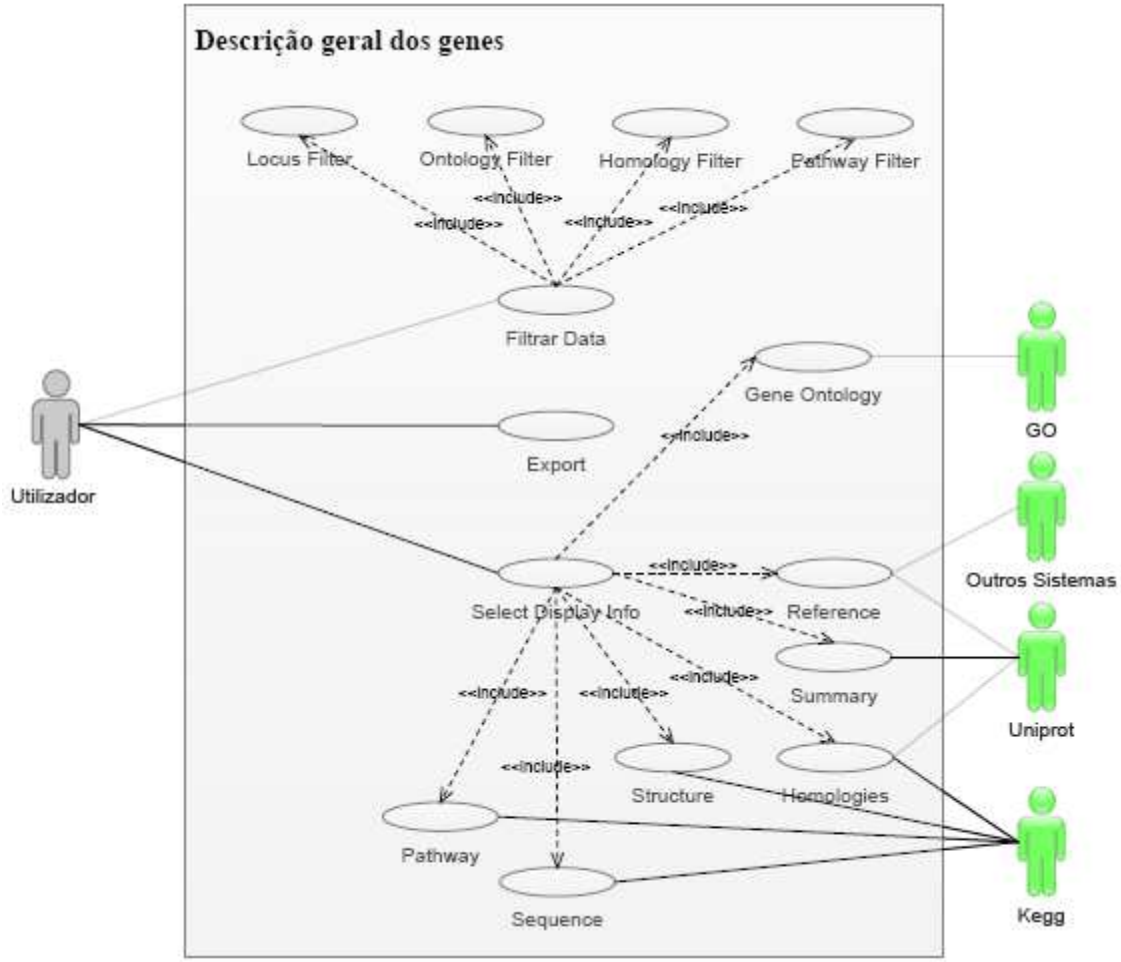



**Figura 3.8 Funcionalidades presentes na descrição geral dos genes.**

## 3.2.4. Descrição ontológica

O grupo de funcionalidades da descrição ontológica (Tabela 3.5 e a Figura 3.9) permite ao utilizador efectuar acções sobre as relações ontológicas que existem entre os genes: selecção de sub ontologia como funções moleculares, componentes celulares e processos biológicos; selecção do tipo de vista em gráfico ou em árvore e ainda permite a selecção dos níveis de ontologia. Toda a informação visível aos utilizadores contém enriquecimento estatístico. Devido à quantidade de informação a ser processada, este processamento é apenas efectuado quando inserimos um *dataset* ou quando alteramos este; após isso a informação é guardada na base de dados.

**Tabela 3.5 Funcionalidades do grupo descrição ontológica.**

| **Grupo** | Funcionalidade | Descrição |
|---|---|---|
| **Descrição ontológica** | Export | Fazer *download* de informação. |
| | Select Display Mode | Permite seleccionar entre vista em árvore e gráfico. |
| | Tree View | Permite a visualização em árvore das classes ontológicas. |
| | Graphic View | Permite a visualização gráfica com a significância das classes de Ontologia. |
| | Select Prune Level | Permite definir o nível dos elementos presentes no gráfico. |
| | Select Sub Ontology | Permite a selecção das subclasses de ontologia. |

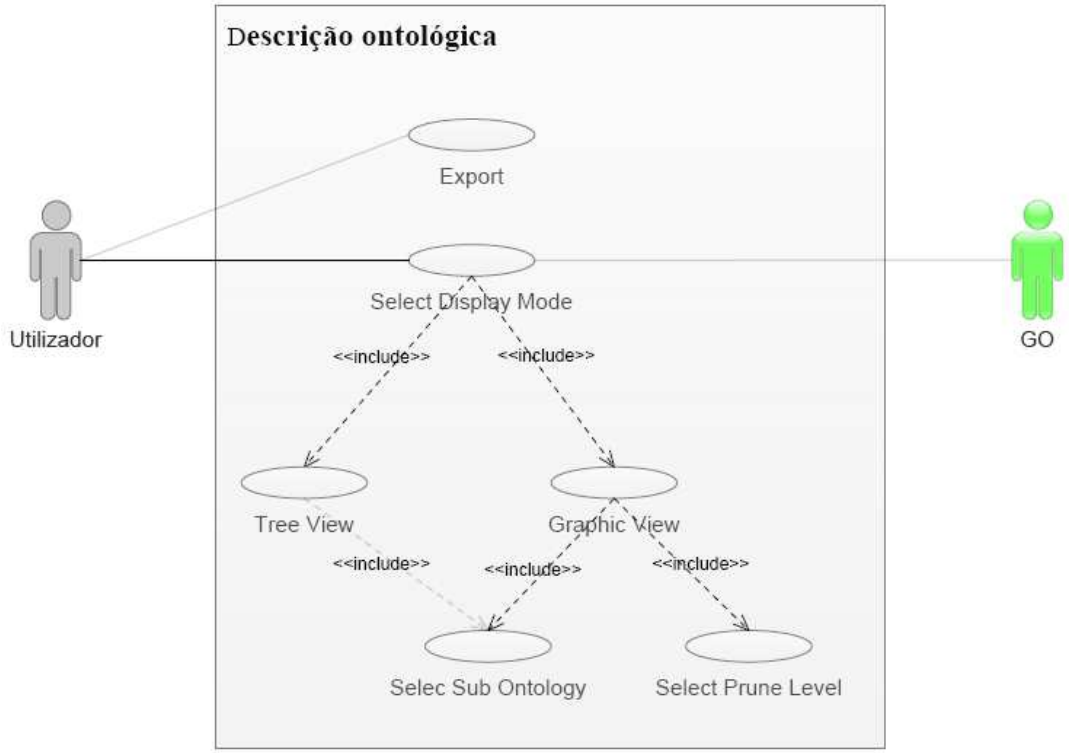



**Figura 3.9 Diagrama de casos de utilização do grupo descrição ontológica.**

### 3.2.5. Homologias

Este grupo de funcionalidades (Tabela 3.6 e Figura 3.10), permite ao utilizador a selecção do tipo de homologia e tipo de fonte de dados. Juntamente com estas opções toda a informação disponibilizada ao utilizador é enriquecida estatisticamente e disponibilizada em forma de gráfico, dando assim uma vista fácil de interpretar, e assim mostrando ao utilizador as classes homológicas mais relevantes para o *dataset* em estudo.

**Tabela 3.6 Funcionalidades do grupo homologias.**

| **Grupo** | Funcionalidade | Descrição |
|---|---|---|
| **Homology** | Export | Fazer *download* de informação. |
| | Select Homology Type | Selecção do tipo de homologia. |
| | Select Database | Selecção da fonte de dados. |
| | Graphic View | Permite a visualização gráfica com a significância das classes de homologia. |

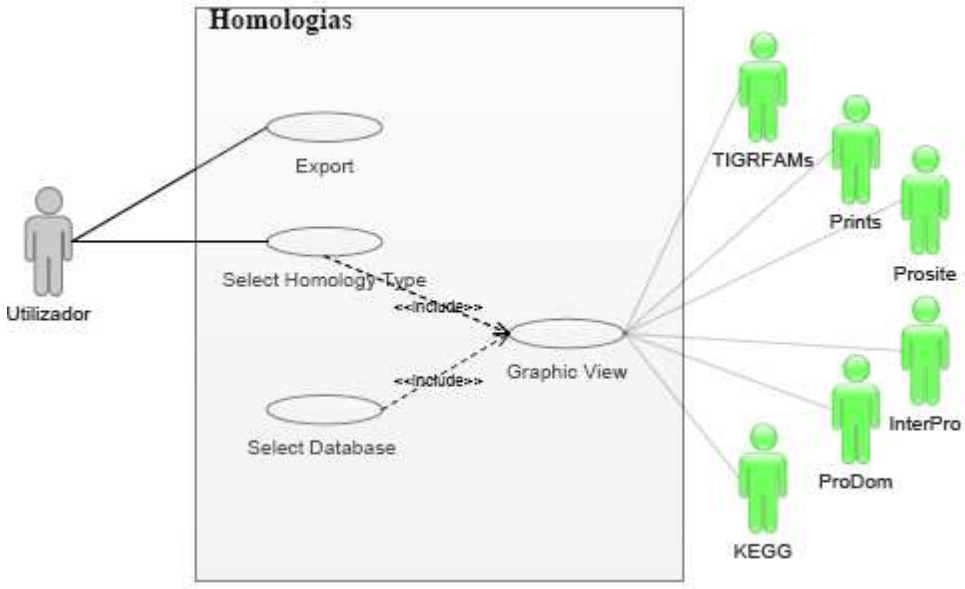



**Figura 3.10 Diagrama de casos de utilização do grupo homologias.**

### 3.2.6. Vias metabólicas

Este grupo de casos de funcionalidades (Tabela 3.7 e Figura 3.11) permite ao utilizador a selecção do tipo de fonte de dados que pretende analisar, uma vista em gráfico, mostrando a relevância estatística de cada via metabólica e o acesso as imagens das vias metabólicas na sua fonte de dados original. O *link* de acesso à fonte de dados tem de conter a anotação dos genes presentes no *dataset* de modo ao utilizador ver estes assinalados na imagem que descreve a via metabólica.

**Tabela 3.7 Funcionalidades do grupo vias metabólicas.**

| **Grupo** | Funcionalidade | Descrição |
|---|---|---|
| **vias metabólicas** | Export | Fazer *download* de informação. |
| | Select Database | Selecção da fonte de dados. |
| | Graphic View | Permite a visualização gráfica com a significância das classes de vias metabólicas. |

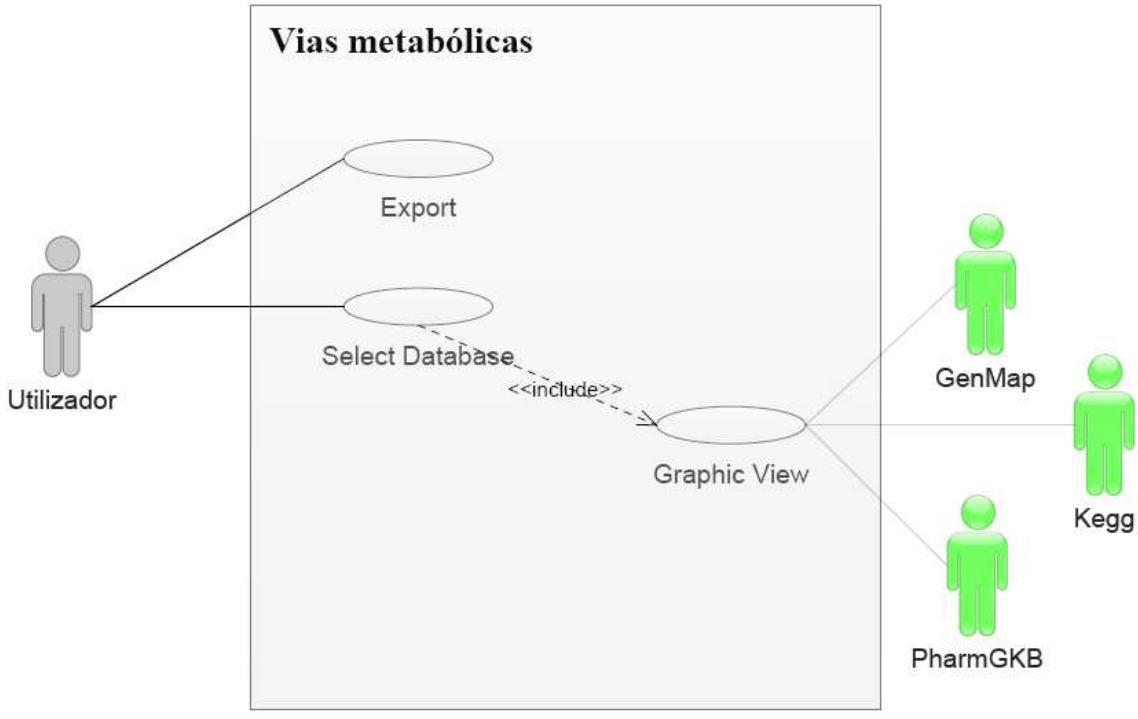



**Figura 3.11 Diagrama de casos de utilização do grupo vias metabólicas.**

### 3.2.7. Localização dos genes nos cromossomas

Este grupo de funcionalidades (Tabela 3.8 e Figura 3.12) deve fornecer ao utilizador:

- Uma vista em gráfico com a anotação dos genes presentes em cada cromossoma, mostrando a relevância de cada cromossoma.
- Link para o NCBI MapViewer com a anotação dos genes presentes em cada cromossoma.

**Tabela 3.8 Funcionalidades do grupo localização dos genes nos cromossomas.**

| Grupo | Funcionalidade | Descrição |
|---|---|---|
| **Localização dos genes nos cromossomas** | Export | Fazer *download* de informação. |
| | Graphic View | |

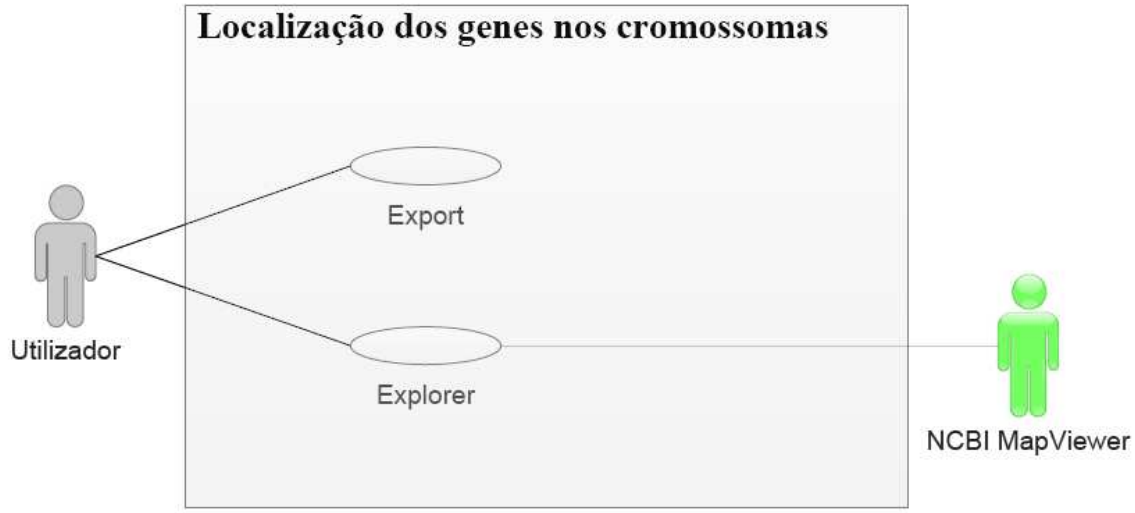



**Figura 3.12 Diagrama de casos de utilização do grupo localização dos genes nos cromossomas.**

### 3.2.8. Dados de experiências anteriores

Este grupo de casos de funcionalidades (Tabela 3.9 e Figura 3.13) permite ao utilizador o acesso ao *ArrayExpress*, disponibilizando assim informação sobre dados de experiências anteriores envolvendo os mesmos genes e os valores da expressão nessas experiências. Podemos ainda visualizar muitos outros dados acedendo a fonte original de informação *ArrayExpress*.

**Tabela 3.9 Funcionalidades do grupo dados de experiências anteriores.**

| **Grupo** | Funcionalidade | Descrição |
|---|---|---|
| **Dados de Experiências anteriores** | Export | Fazer *download* de informação. |
| | Explorer | Permite a navegação do utilizador pelos dados processados. |


**Figura 3.13 Diagrama de casos de utilização do grupo dados de experiências anteriores.**

### 3.2.9. Bibliografia

Este grupo de casos de funcionalidades (Tabela 3.10 e Figura 3.14) permite ao utilizador a visualização dos artigos mais relevantes nos quais o seu grupo de genes é referenciado, permite ainda o acesso ao NCBI Pubmed. Devido à quantidade de dados a processar, estes dados apenas são processados quando um *dataset* é inserido ou alterado de modo a diminuir o tempo de acesso aos dados.

**Tabela 3.10 Funcionalidades do grupo Bibliografia.**

| Grupo | Funcionalidade | Descrição |
|---|---|---|
| **Bibliografia** | Export | Fazer download de informação. |
| | Explorer | Permite a navegação do utilizador pelos dados processados. |

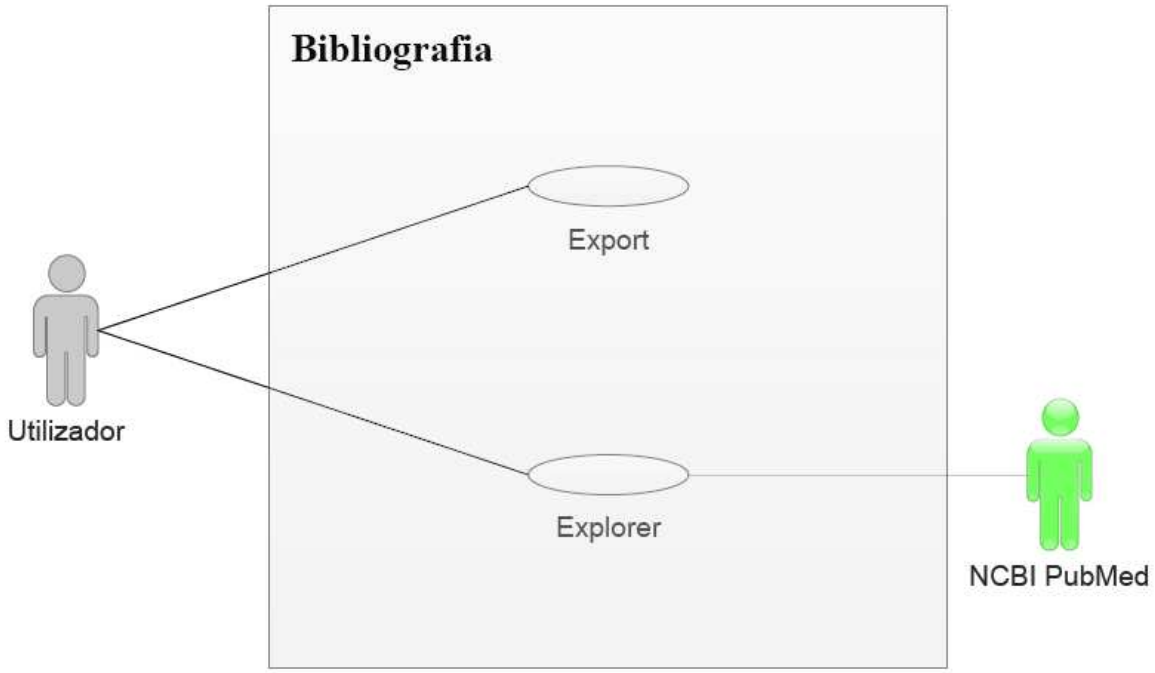



**Figura 3.14 Diagrama de casos de utilização do grupo bibliografia.**

## 3.3. Requisitos Técnicos

Além dos requisitos funcionais, para a aplicação ter as características que desejamos necessita de cumprir diversos requisitos de nível técnico (não funcionais). É a partir destes requisitos que será estruturado o sistema de suporte à aplicação e às funcionalidades da mesma.

Relativamente ao interface este deve ser flexível e permitir uma actualização de forma independente do resto da aplicação. Além disso, deve ainda usar fontes e cores que facilitem a legibilidade da informação, e conter gráficos facilmente analisáveis para facilitar a interpretação dos dados.

Quanto ao desempenho o sistema, deve-se garantir a rápida inserção e validação de dados bem como a consulta rápida de informação.

Os requisitos de segurança implicam que os dados de utilizadores registados sejam permanentes e privados. Para os utilizadores não registados o sistema deve manter os dados privados e temporariamente disponíveis de modo a que este utilizador os possa consultar.

Quanto a requisitos de interface com sistemas externos e com ambientes de execução, todas as aquisições de dados são feitos em sistemas externos como o GeNS (*Gene Names Server*), GO [8], NCBI-Pubmed [10]. Devido aos requisitos de desempenho alguns destes dados são processados e guardados na base de dados do sistema.

Quanto a especificações genéricas, o sistema foi construído, de forma a ser compatível com a generalidade de *browsers* existentes, principalmente os mais usados como *Firefox*, *Internet Explorer*, *Safari* e *Opera*.

Deve ser fácil actualizar os componentes construídos, ou adicionar novos componentes. Os componentes devem desta forma ser independentes entre si, de modo a facilitar a actualização.

## 3.4. Métodos Estatísticos

Quando se analiza dados, deseja-se a conclusão mais forte para uma quantidade limitada de dados. Para fazer isso temos de ultrapassar dois problemas:

- A obscuridade de importantes diferenças na variabilidade biológica e imprecisão experimental, dificultam a distinção entre diferenças reais e variabilidade aleatória.
- A tendência de procura de padrões, mesmo em dados aleatórios, que o nosso cérebro tem. A nossa inclinação natural (especialmente com os nossos dados) é concluir que as diferenças são reais e minimizar a contribuição da variabilidade aleatória. O rigor dos métodos estatísticos previne de cometermos este tipo de erros.

A análise estatística é muito útil quando estamos a procura de diferenças que são pequenas comparadas com a imprecisão experimental e variabilidade biológica. Se apenas nos preocupamos com diferenças grandes, podemos encontrar essas diferenças directamente nos dados.

### 3.4.1. População vs. Amostra

A lógica presente nos métodos estatísticos assume que a amostra é seleccionada aleatóriamente de uma população. Isto funciona bem para controlo de qualidade, mas quando se aplica isto a dados científicos, encontram-se dois problemas:

- Não existe uma amostra aleatória da população.
- Resultam conclusões que vão para além da população.

Raramente se selecciona uma amostra aleatória de uma população, usualmente só se faz uma experiência algumas vezes para explorar a situação global.

No nosso caso, a amostra é a lista de genes diferencialmente expressos (GeneList1) e a população todos os genes da experiência ou organismo (GeneList2).

### 3.4.2. Independência

Não é suficiente os nossos dados serem uma amostra de uma população. Os testes estatísticos são também baseados na presunção que cada unidade da experiência foi tirada de forma independente.

Por exemplo quando se faz uma experiência de laboratório três vezes, cada vez em triplicado, apenas se obtém três amostras independentes, que são as médias dos três valores medidos para cada experiência.

### 3.4.3. P-value

Considerando uma experiência onde são medidos valores de duas amostras e as médias são diferentes. Com quanta certeza se pode dizer que a média da população é diferente? Existem duas possibilidades:

- A população tem médias diferentes.
- A população tem a mesma média e a diferença observada é fruto de uma coincidência de amostras aleatórias.

O p-value é uma probabilidade, com um valor entre 0 e 1, e permite responder a esta questão: Se a população tem realmente a mesma média, é provável que uma amostra aleatória leve a uma diferença entre as médias tão grande ou maior que a observada.

Existem diferentes métodos de cálculo do p-value como o *binomial*, *chi-square*, *Fisher's* e *hipergeométrico*. A escolha do teste estatístico depende de como são expressos os dados de uma experiência (medidas, tempo, proporção, etc), se podemos tratar estes agrupados e assumindo que os valores medidos seguem uma distribuição *Gaussiana* [67].

**Hipergeométrica**

Tendo N bolas (genes) nos quais M são vermelhos e N-M são verdes, escolhendo K bolas e perguntando qual é a probabilidade de se ter escolhido x bolas vermelhas.

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}$$

Tendo em conta o definido acima, a probabilidade de termos x ou menos genes em F pode ser calculada pelo somatório das probabilidades.

$$pu = P((X = 1) + P(X = 2) + \cdots + P(X = x)) = \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$$

Isto corresponde a um teste para o qual um p-value pequeno corresponde a uma categoria pouco representada. Então para efectuarmos a normalização do p-value obtemos:

$$p0 = 1 - pu = 1 - \sum_{i=0}^{x} \frac{\binom{M}{i}\binom{N-M}{K-i}}{\binom{N}{K}}$$

Temos de ter atenção com o método hipergeométrico pois é difícil de calcular quando o número de genes é muito elevado. Contudo quando N é muito elevado a distribuição hipergeométrica tende para a binomial.

**Binomial**

Se for usada a distribuição binomial, a probabilidade de se obter x genes em F num conjunto de K genes seleccionados aleatoriamente é dado pela forma clássica da binomial.

$$P(X = x|K, \frac{M}{N}) = \binom{K}{x}\left(\frac{M}{N}\right)^{x}\left(1 - \frac{M}{N}\right)^{K-x}$$

E o p-value é dado por:

$$p = \sum_{i=0}^{x}\binom{K}{i}\left(\frac{M}{N}\right)^{i}\left(1 - \frac{M}{N}\right)^{K-i}$$

## Chi-square

**Tabela 3.11 Significância das categorias funcionais em F podem ser calculadas usando uma matriz 2x2 e os métodos chi-square ou Fisher's [67].**

| | Genes no Array | Genes Diferencialmente expressos | |
|---|---|---|---|
| Estão Contidos | $n_{11}$ | $n_{12}$ | N1.=$\sum_{j=1}^{2} n_{1j}$ |
| Não estão contidos | $n_{21}$ | $n_{22}$ | N2.=$\sum_{j=1}^{2} n_{2j}$ |
| | N.1=$\sum_{i=1}^{2} n_{i1}$ | N.2=$\sum_{i=1}^{2} n_{i2}$ | N..=$\sum_{i,j} n_{ij}$ |

Com esta noção o número de genes do *microarray* é N=N.1, o número de genes na categoria funcional F é M=$n_{11}$, o número de genes diferencialmente expressos K=N.2 e o número de genes diferencialmente expressos de F x=$n_{12}$. Usando esta noção, podemos calcular o $X^2$:

$$X^2 = \frac{N..(|n_{11}n_{22} - n_{12}n_{21}| - \frac{N..}{2})^2}{N1.N2.N.1N.2}$$

No entanto o cálculo do $X^2$ não pode ser usado para amostras pequenas. A regra é que para o teste dar uma conclusão válida temos de ter $E_{ij} = \frac{N_{i.} \cdot N_{.j}}{N_{..}}$ >5.

## Fisher's

O teste exacto de *Fisher's* considera todas as linhas e colunas da matriz (Tabela 3.11) fixas e usa a distribuição hipergeométrica para calcular a probabilidade de observação de cada combinação independente da tabela:

$$P = \frac{N1.!\,N2.!\,N.1!\,N.2!}{N..!\,n_{11}!\,n_{12}!\,n_{21}!\,n_{22}!}$$

O p-value de uma ocorrência particular é calculado com a soma de todas as probabilidades menores que a probabilidade de observação:

A probabilidade de observação é dada por:

$$P(n12|n11) = \left(\frac{N.2}{N.1}\right)^{n12} . \frac{(n11+n12)!}{n11!.n12!.\left(\frac{N.2}{N.1}\right)^{n11+n12+1}}$$

E o p-value é:

$$p = \min\left\{\sum_{k=0}^{k\leq n12} P(k|n11), \sum_{k=n12}^{\infty} P(k|n11),\right\}$$

Olhando para um exemplo se o valor de p-value é 0.03, o que se pode concluir que uma amostra aleatória da mesma população levaria a uma diferença menor que a observada 97% das vezes e a uma diferença maior que a observada 3% das vezes.

### 3.4.4. P-value Corrigido

No contexto do trabalho desenvolvido o p-value é a probabilidade de um gene ter uma discrepância aleatória entre as classes que o representam, se não existir essa discrepância.

O p-value corrigido pelo método de *Bonferroni* é a multiplicação do p-value pelo número de genes da experiência e é igual ao número de genes que esperávamos que tivessem um p-value tão pequeno aleatoriamente dado o tamanho da experiência. Por exemplo, numa experiência com 1000 genes, era esperado ter um p-value menor que 0,001. O FDR (*False Discovery Rate*) é a fracção de genes iguais ou inferiores a um determinado p-value que se prevê que venha a ter esses pequenos valores: é numericamente igual ao p-value corrigido de *Bonferroni* dividido pelo número de genes com o p-value menor.

### 3.4.5. Intervalos de confiança

A significância estatística pode ser definida em termos de intervalos de confiança. Mas a consideração de um valor de p-value como estatisticamente significante é um pouco mais complexa. Geralmente o valor considerado é 95%, no entanto não há nada de especial em 95%, é apenas convenção que é geralmente usada no cálculo dos intervalos de confiança. Em teoria, os intervalos de confiança podem ser calculados para qualquer grau de confiança. Se se pretender maior confiança, os intervalos serão mais amplos, se está disposto a aceitar uma menor confiança, os intervalos serão reduzidas [68].

## 3.5. Arquitectura da Aplicação

Numa aplicação *Web*, mais que nas tradicionais aplicações *Desktop*, os recursos necessários para o bom funcionamento da aplicação são variados. Tendo em conta que a aplicação precisa de ser escalável, flexível e estar *online* 24h por dia, torna-se essencial a utilização de diversos componentes para que o serviço seja mais completo.

### 3.5.1. Componentes

Antes de começar o desenvolvimento, a aplicação teve de ser planificada, sendo escolhidas as tecnologias que seriam usadas para responder aos objectivos propostos, cumprindo os requisitos traçados. As escolhas foram feitas para que a aplicação se integrasse da melhor forma possível com o conjunto de aplicações do grupo já existentes, simplificando a selecção de alguns dos componentes.

***Páginas Web Dinâmicas***

Existem várias *frameworks* de desenvolvimento de páginas *Web* dinâmicas sendo as mais conhecidas ASP, PHP, JSP e mais recentemente *Ruby on Rails*.

Foi decidido usar no desenvolvimento desta aplicação ASP (*Active Server Pages*), decisão tomada devido ao *GeneBrowser* ter de interagir com outra aplicação (GeNS) desenvolvida em tecnologias *Microsoft*. O GeNS, está a ser desenvolvido com tecnologia *SQL Server* da *Microsoft*, e devido à

existência de grande compatibilidade entre as tecnologias *Microsoft* a melhor solução encontrada foi ASP.

O ASP é uma *framework* em que se podem usar *CSharpScript*, *VBScript*, *JScript*, *PerlScript* ou *Python* processadas pelo lado servidor para geração de conteúdo *Web* dinâmicos. A linguagem de script é interpretada do lado do servidor, e o que é enviado para o cliente é apenas a saída, que normalmente é uma linguagem de marcação como HTML, XHTML ou XML e também *JavaScript*.

***Servidor Web***

Após definirmos a utilização de ASP.Net é fácil de perceber que o servidor *Web* mais indicado para este tipo de aplicação é o IIS. O IIS (*Internet Information Services*) é um servidor *Web* criado pela *Microsoft* para seus sistemas operativos. A sua primeira versão foi introduzida com o *Windows NT Server versão 4* e passou por várias actualizações até a versão 7, que irá ser usada neste projecto.

***Sistema de Gestão de Bases de Dados (SGBD)***

Quanto ao SGBD os mais conhecidos são: *Oracle*, *Microsoft SQL Server*, IBM DB2, MySql entre vários outros. O objectivo principal de um SGBD é prover um ambiente que seja adequado e eficiente para uso na recuperação e armazenamento de informações.

Algumas funções extremamente relevantes do SGBD, são:

- Interacção com o sistema de arquivos do sistema operacional.
- Cumprimento da integridade.
- Cumprimento da segurança.
- Cópias de segurança (“backup”) e recuperação.
- Controle de concorrência.
- Optimização e execução dos comandos DML.
- Dicionário de Dados.
- Desempenho

Esta foi a que se decidiu utilizar neste trabalho, devido a compatibilidade existente entre as tecnologias *Microsoft* foi *SQL Server*. O *SQL Server* é o principal concorrente do *Oracle* e tem como vantagem o facto de ser da *Microsoft* e se integrar nativamente com os seus produtos e linguagens. Actualmente o *SQL Server* conta também com uma boa participação no mercado de

*Web*, fruto de um relacionamento mais estreito com as linguagens ASP e ASP.NET que lideram o mercado de médios e grandes projectos de *Internet*.

### 3.5.2. Modelo

O modelo de suporte à aplicação revela-se simples e objectivo. Mesmo tendo um extenso número de requisitos a cumprir e recursos a usar, o modelo traduz-se em quatro camadas distintas (Figura 3.15).

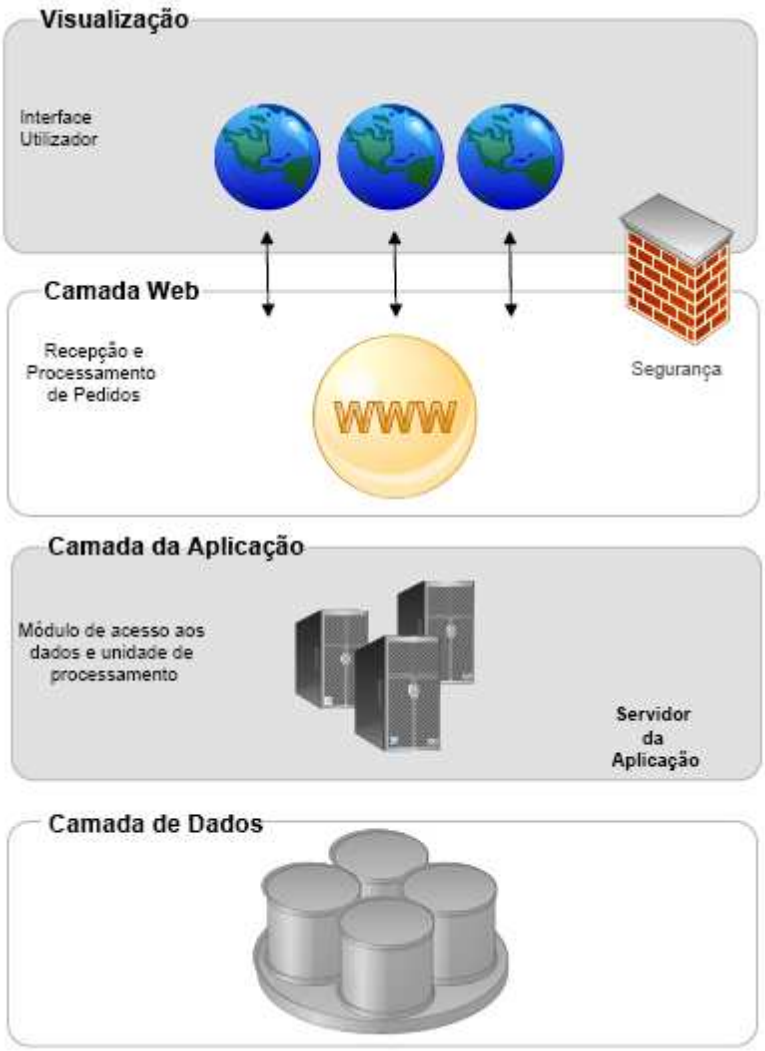



**Figura 3.15 Modelo de componentes da solução.**

Começando pelo nível inferior, tem-se a primeira camada, de acesso aos adaptadores e ao SGBD da aplicação. Acima desta tem-se a Camada da Aplicação. Nesta camada, apresentam-se os métodos de acesso aos dados e os módulos de processamento do lado do servidor, onde é tratado o conteúdo das páginas a mostrar ao utilizador e a visualização da resposta a algumas operações. De seguida tem-se a camada *Web* que é responsável pela recepção e processamento de pedidos do cliente. No topo, visualiza-se a camada do cliente, onde se encontra o *browser* que o utilizador usa para aceder à aplicação.

### 3.5.3. Arquitectura e Instalação

A arquitectura da aplicação é mais complexa que o modelo. Sucintamente, o funcionamento da aplicação subdivide-se em três categorias. Acessos do cliente ao *Servidor Web*: quando o utilizador abre o site e utiliza a aplicação, os pedidos são efectuados directamente ao *Servidor Web* e tratados por este. Acessos do *Servidor Web* aos serviços externos: caso seja necessário para a realização de uma determinada funcionalidade e o *Servidor Web* tem acesso ao Servidor de Base de Dados.

Um possível diagrama de instalação da aplicação é o presente na Figura 3.16, um servidor de base de dados, pode ou não ser independente do resto do sistema e o *Servidor Web* que é o elo de ligação entre o sistema e o exterior.

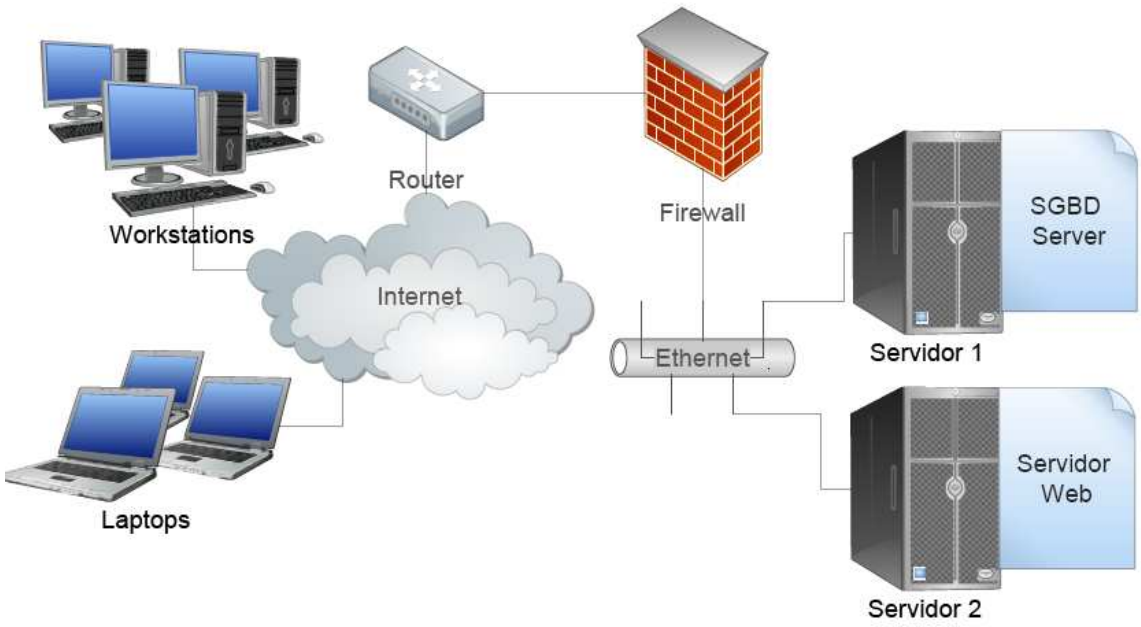



**Figura 3.16 Diagrama de instalação do *GeneBrowser*.**

## 3.6. Sumário

Neste capítulo foi apresentada uma proposta de arquitectura do *GeneBrowser*, este é apresentado como uma aplicação multi-camada, que faz a integração de dados e serviços de múltiplas fontes, tendo como fonte principal de dados o GeNS.

O *GeneBrowser* é apresentado como uma plataforma que incorpora as vantagens dos sistemas que fazem integração de dados biológicos com as dos sistemas que fazem enriquecimento estatístico. Para isso foi apresentado a proposta de desenvolvimento que tem em vista o enriquecimento estatístico das classes funcionais que descrevem os genes.

# Capítulo 4 - GeneBrowser 2.0

## 4.1. Estratégias de Desenvolvimento

A estratégia para levar a cabo este trabalho, inicialmente consistiu em, aprofundar o conhecimento sobre o sistema implementado. O sistema foi implementado em *Visual Studio Net* sendo a base de dados *SQL Server 2008*. O sistema foi todo desenvolvido de forma a potencializar um bom desempenho em todas as funcionalidades presentes. Para tal não foram usados componentes *.Net*. Antes foram desenvolvidos componentes para geração de gráficos, árvores e *gridviews*. Estes componentes foram desenvolvidos de modo a receberem os dados em JSON e gerar os componentes em HTML, diminuindo assim a quantidade de dados a ser transmitida aos clientes, bem como o processamento do lado do servidor.

### 4.1.1. Processamento e métodos estatísticos

Foi desenvolvida uma classe de cálculo estatístico que contém os diferentes métodos de cálculo de p-value e um método de cálculo p-value corrigido (Figura 4.1).

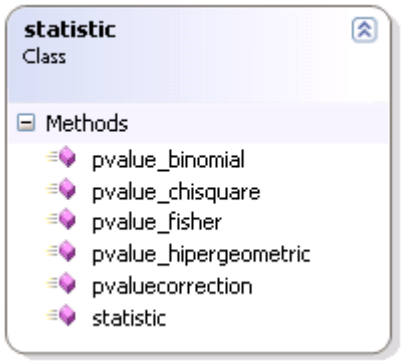



**Figura 4.1 Classe com métodos estatísticos.**

Devido à quantidade de informação presente em algumas funcionalidades, nomeadamente em *Gene Ontology* e *Bibliography*, foi detectado que devido à quantidade de informação a processar, eram perdidos os pedidos dos outros clientes no IIS. De modo a solucionar este problema foi

desenvolvida uma aplicação que executa o cálculo e processamento destas informações, guardando os dados processados na base de dados do *GeneBrowser*.

Esta aplicação tem interligação com o *GeneBrowser* e sempre que é inserido um novo dataset, esta é lançada iniciando o processamento dos dados. Deste modo, o *GeneBrowser* sabe quando o processamento é iniciado, e para saber quando este é concluído, foi criado um campo na base de dados na tabela dataset que indica o estado do processamento. Na Figura 4.2 observa-se o diagrama de classes desta aplicação.

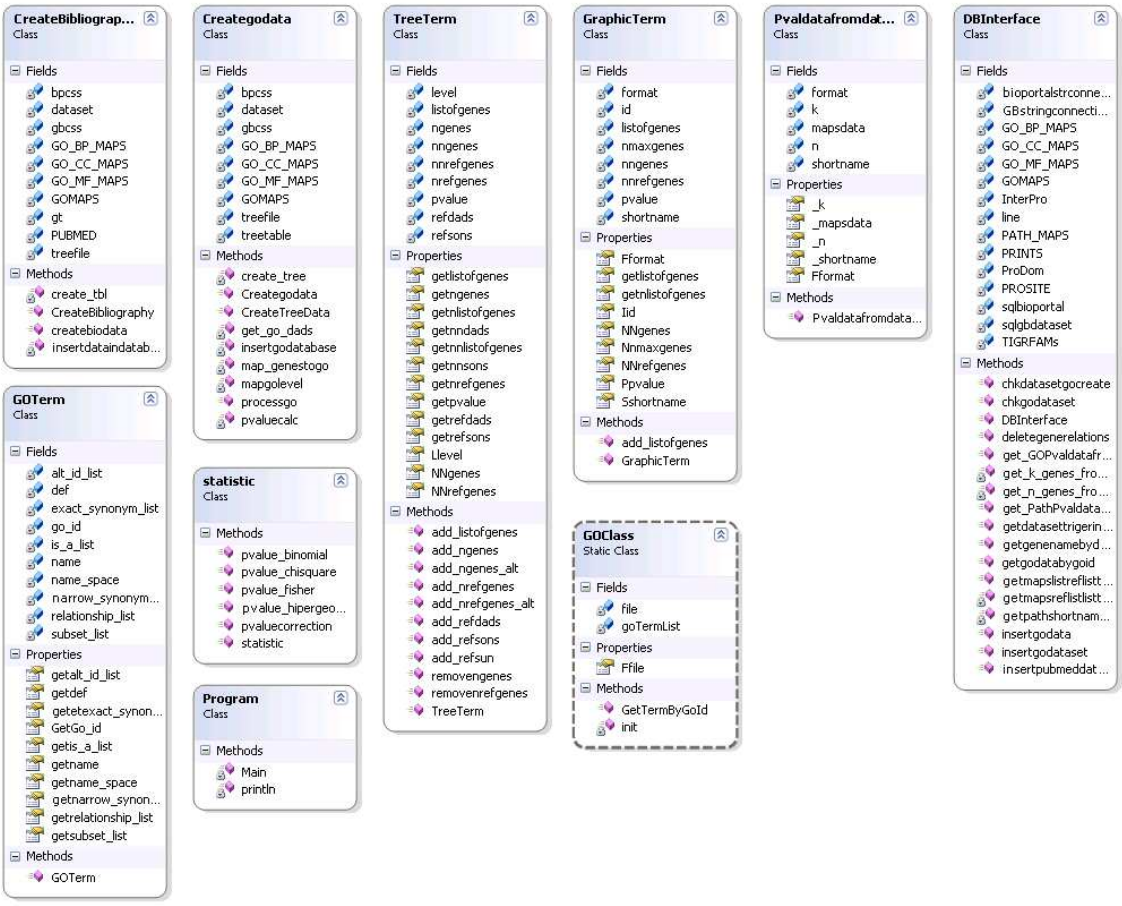



**Figura 4.2 Diagrama de classes da aplicação de processamento.**

## 4.1.2. Gráficos

As notações gráficas existem há algum tempo e o seu principal valor está na comunicação e no entendimento. Um bom diagrama frequentemente pode ajudar a transmitir ideias sobre um projecto, particularmente quando se pretendem evitar muitos detalhes.

Existem muitas bibliotecas que geram gráficos, no entanto estas inserem muito conteúdo indesejado nas páginas html. Devido a esse facto, e por já termos muita informação a ser transmitida, optou-se por desenvolver a uma solução própria de geração de gráficos.

Esta biblioteca é composta por três componentes:

a) A estrutura de dados em JSON.
b) O estilo definido em CSS.
c) A geração de gráficos em Javascript.

Os gráficos gerados são do género do apresentado na Figura 4.3.

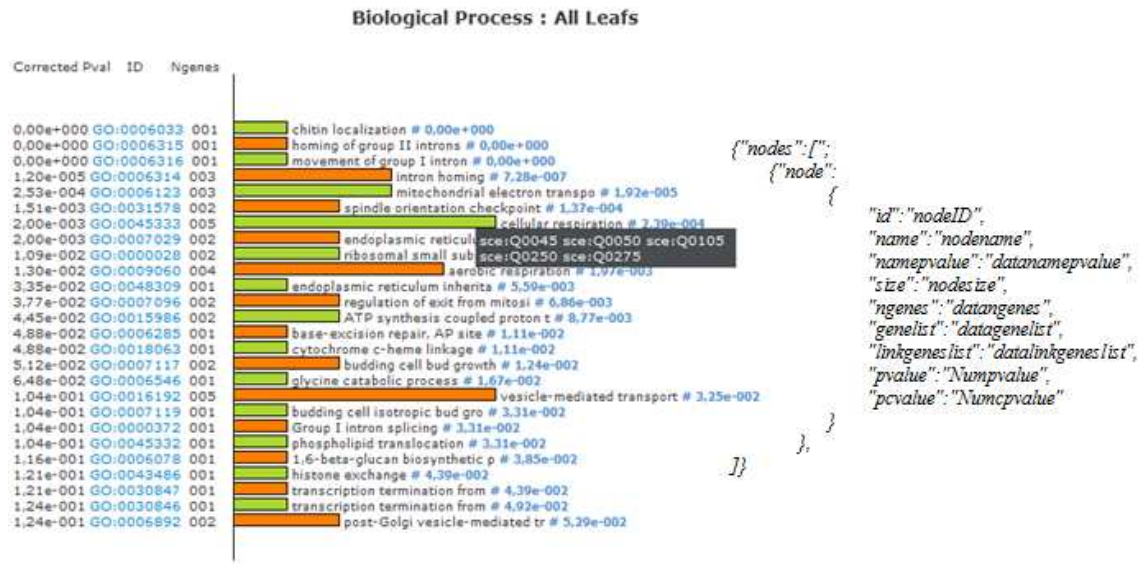



**Figura 4.3 Exemplo de um gráfico, e a estrutura de dados em JSON.**

Antes do processo de criação de um gráfico, é necessário realizar várias operações do lado do servidor, nomeadamente a obtenção dos dados, cálculo do p-value, cálculo do p-value corrigido e a conversão dos dados para JSON. Estas operações geralmente exigem grande quantidade de processamento e volume de dados, pelo que se optou, nas funcionalidades que contém maior volume de dados, efectuar o processamento numa aplicação separada. Após efectuar estes passos os dados são transmitidos ao cliente e o processamento responsável pela geração do gráfico é efectuado do lado do cliente.

### 4.1.3. Árvores

A mesma abordagem que nos levou a construir uma biblioteca para geração e gráficos, levou-nos a construção de componentes para geração de uma estrutura em árvore.

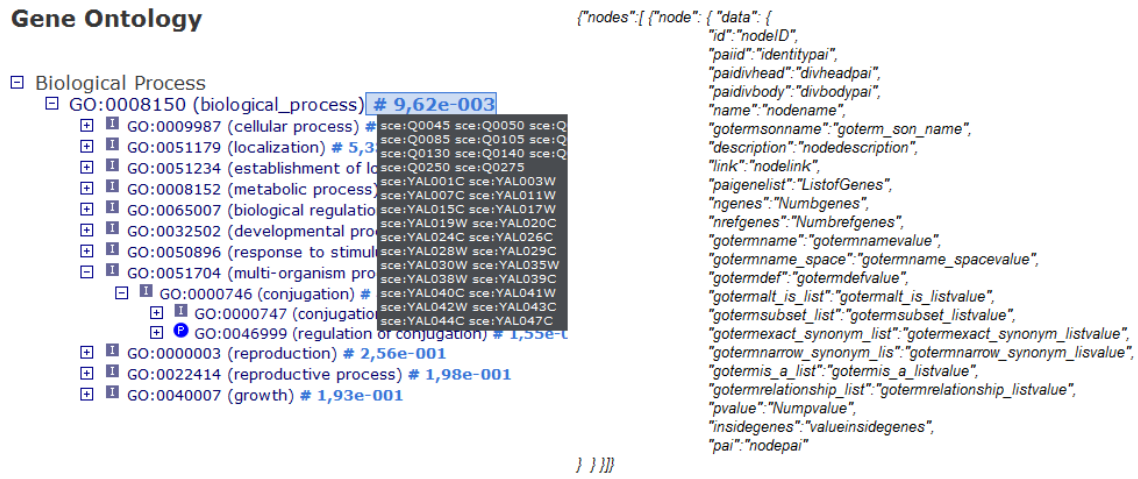



**Figura 4.4 Exemplo de árvore e estrutura de dados em JSON.**

Na Figura 4.4 visualiza-se a estrutura da árvore, bem como a estrutura de dados em JSON que leva à construção da mesma. A informação necessária à construção da árvore é processada e guardada na base de dados, enquanto a árvore é construída dinamicamente com *Javascript* e JSON.

Quando se clica num elemento da árvore é feito um pedido ao servidor, o que responde com uma estrutura de dados em JSON, informação que é processada no cliente e que leva à construção da árvore, elemento por elemento.

### 4.1.4. Gridviews

Para a apresentação de tabelas (*gridviews*) optou-se por seguir a mesma abordagem dos componentes anteriores, como se pode ver na Figura 4.5

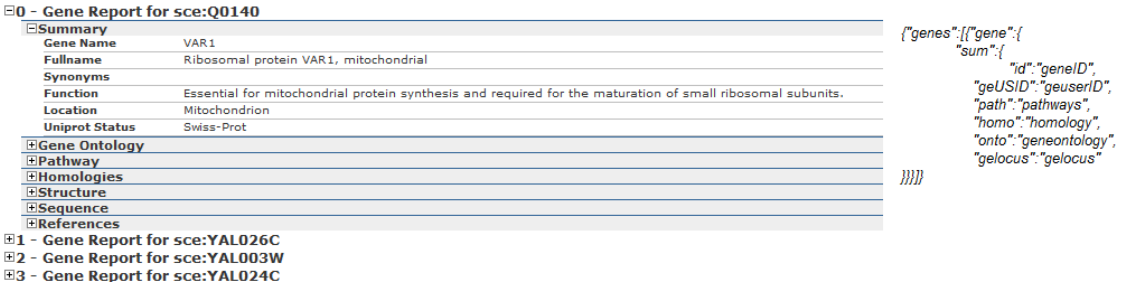



**Figura 4.5 Exemplo do gridview e a estrutura de dados em JSON.**

# 4.2. Arquitectura

## 4.2.1. Diagrama de base de dados da aplicação

Para suportar o armazenamento da informação dos *Dataset*, bem como a informação referente a cada utilizador, foi necessário elaborar um esquema de base de dados seguindo o modelo relacional, que guardasse estes dados e permitisse um desempenho óptimo na execução das funcionalidades da aplicação.

A Figura 4.6 esquematiza a estrutura da base de dados, mostrando as diversas tabelas criadas e os tipos de dados de cada campo.

Como se pode comprovar, o esquema da base de dados é relativamente simples. A Tabela 4.1 mostra uma pequena descrição dos elementos pertencentes a cada tabela.

## 4.2.2. Diagrama de classes da aplicação

A organização por classes da aplicação encontra-se esquematizada na Figura 4.7.

De acordo com o modelado, o acesso à base de dados realiza-se numa camada única sintetizada na classe DBInterface. É nesta classe que são implementados todos os métodos de comunicação com as bases de dados. Todas as classes de processamento massivo de informação são independentes entre si e independentes da recepção de pedidos feitos pelo cliente. Esta estrutura tenta pôr em

prática, da forma mais coerente possível, o modelo proposto. O pré-processamento das páginas não usa a classe DBInterface directamente, mas sim uma instanciação da mesma presente em cada uma das classes de processamento da aplicação.

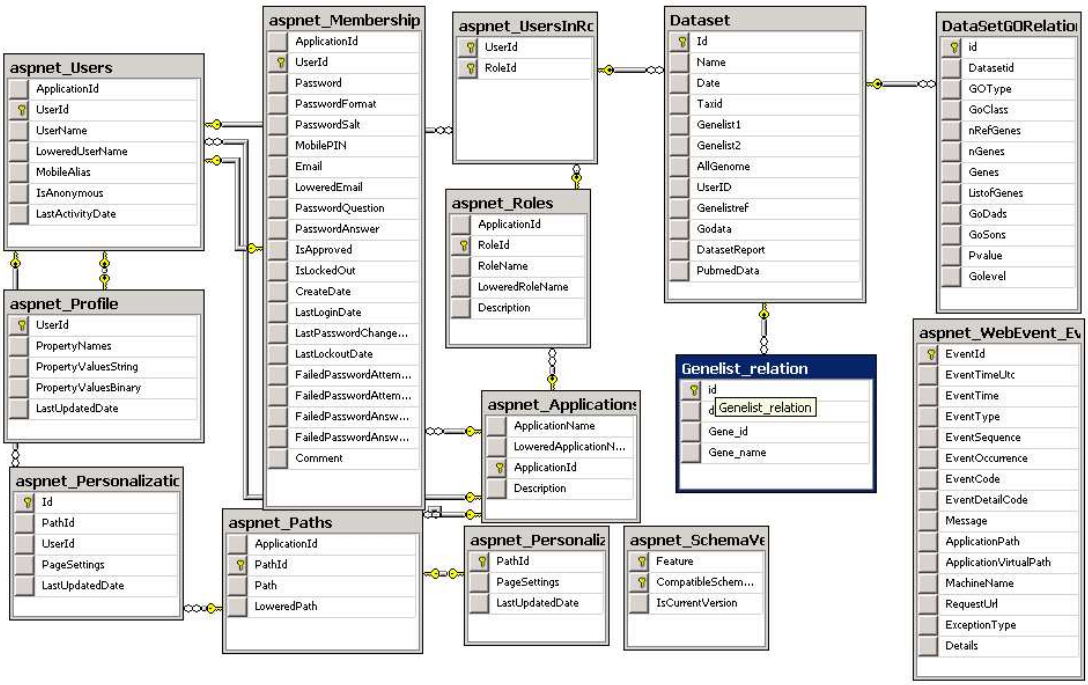



**Figura 4.6 Diagrama de classe das bases de dados.**

**Tabela 4.1 Descrição das relações da base de dados**

| Grupo | Campo | Descrição |
|---|---|---|
| **Dataset** | **id** | Identificador do *dataset*. |
| | **Name** | Nome do *dataset* |
| | **Date** | Data de criação do *dataset* |
| | **Taxid** | Identificador de espécie |
| | **GeneList2** | Lista de genes total do *dataset*. |
| | **AllGenome** | Selecção de todo o genoma ou lista de genes 2. |
| | **UserID** | Identificador do utilizador. |
| | **Processing** | Sinalização do processamento de dados. |
| | **DatasetReport** | Relatório de inserção de dados. |
| **Genelist_relation** | **id** | Identificador de relação gene *dataset*. |
| | **DatasetId** | Identificador de dataset. |
| | **Geneid** | Identificador de gene. |
| | **GeneName** | Nome do gene introduzido pelo utilizador. |
| **DataSetProcRelation** | **Id** | Identificador de relação dados processada/*dataset*. |
| | **Datasetid** | Identificador de *dataset*. |
| | **Type** | Tipo de dados a Guardar. |
| | **Class** | Identificador dos dados a guardar. |
| | **NrefGenes** | Número de genes de referência. |
| | **NGenes** | Número de genes. |
| | **Genes** | Genes presentes neste elemento. |
| | **ListofGenes** | Lista de genes associados a este elemento. |
| | **Dads** | Pais esclusivo para dados de ontologias. |
| | **Sons** | Filhos esclusivo para dados de ontologias. |
| | **P-value** | Valor do p-value. |
| | **Level** | Nível da árvore esclusivo para dados de ontologias. |

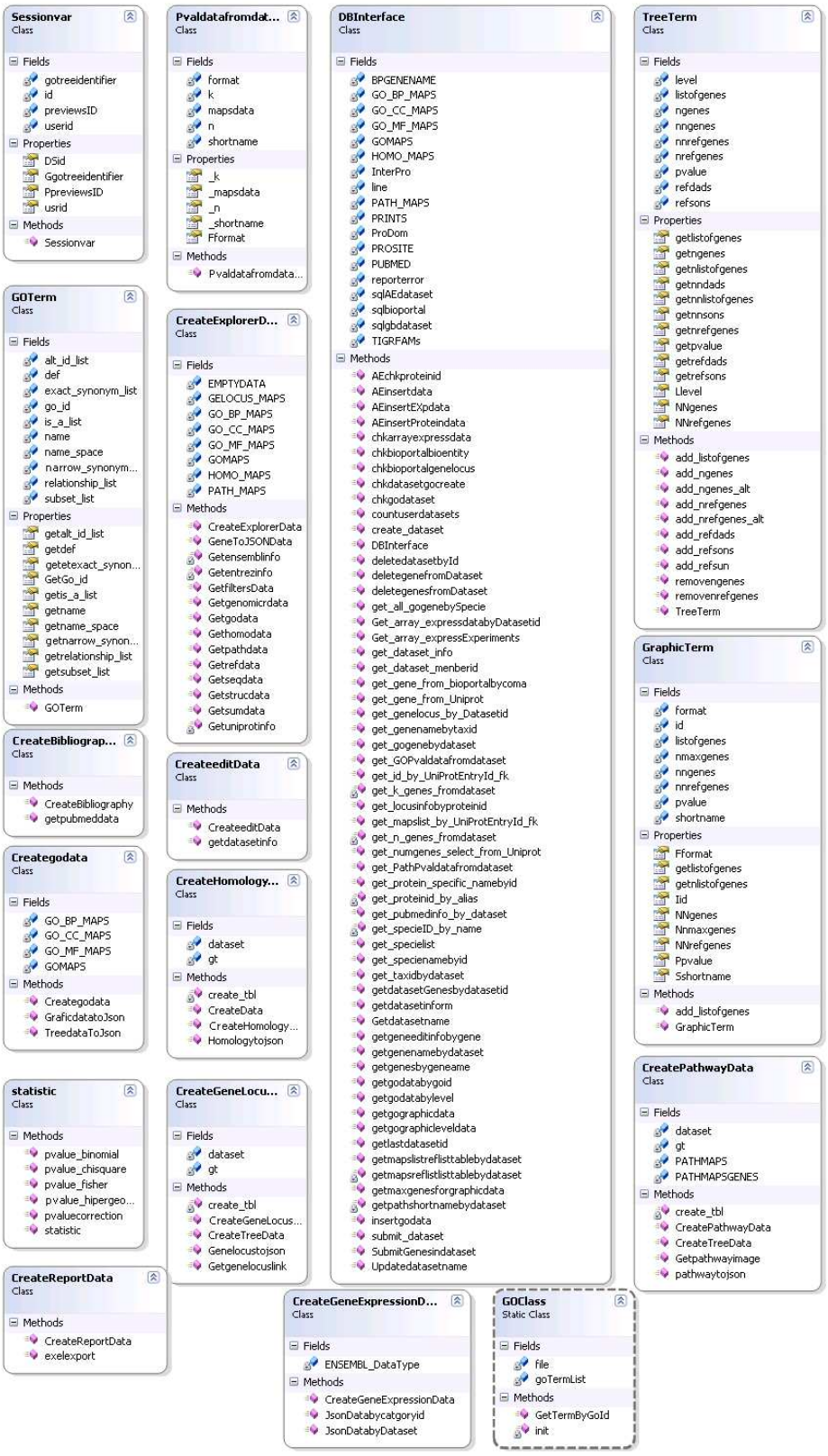



**Figura 4.7 Diagrama de classes da aplicação.**

## 4.3. Interface de Utilizador

O *GeneBrowser* é uma aplicação *Web* disponível em http://bioinformatics.ua.pt/GeneBrowser2/. A sua utilização não requer qualquer tipo de autenticação, no entanto, de modo a manter um histórico de utilização guardando as experiências previamente inseridas, o utilizador terá de se registar no sistema. Após ter alcançado o *Website*, o utilizador pode registar-se, ver *demos* de funcionamento do sistema e ter ajuda na inserção de uma nova experiência.

Com o *GeneBrowser* o utilizador pode obter enriquecimento funcional de genes utilizando várias definições (ontologias, homologias, Ortologias, vias metabólicas, etc), informações sobre a expressão génica em estudos anteriores (*ArrayExpress*) e as mais relevantes publicações (*PubMed*).

Na Figura 4.8 está presente o esquema de área de trabalho principal do *GeneBrowser*.


**Figura 4.8 Esquema da área de trabalho.**

### 4.3.1. Home

Quando o utilizador chega à página inicial do *GeneBrowser*, depara-se com várias opções: inserção de um novo *dataset*, registar-se ou se já for utilizador registado pode entrar no sistema (Figura 4.9).

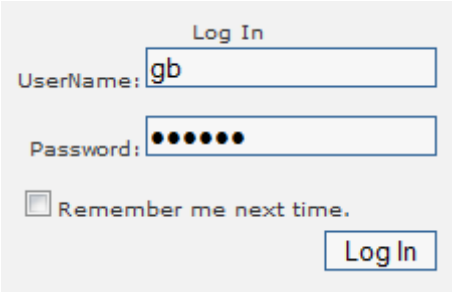



**Figura 4.9 Login no *GeneBrowser*.**

O *GeneBrowser* permite a análise de genes de 835 organismos e permite a inserção de 27 tipos diferentes de identificadores de genes/proteínas. Algum destes identificadores são de bases de dados específicas de organismo (contém apenas um organismo) como o *Hugo* e CGD, as outras são de bases de dados genéricas (contém mais de um organismo) como *KEGG*, NCBI, *Uniprot*. Num mesmo *dataset* o utilizador pode inserir diversos tipos de identificador de genes/proteínas.

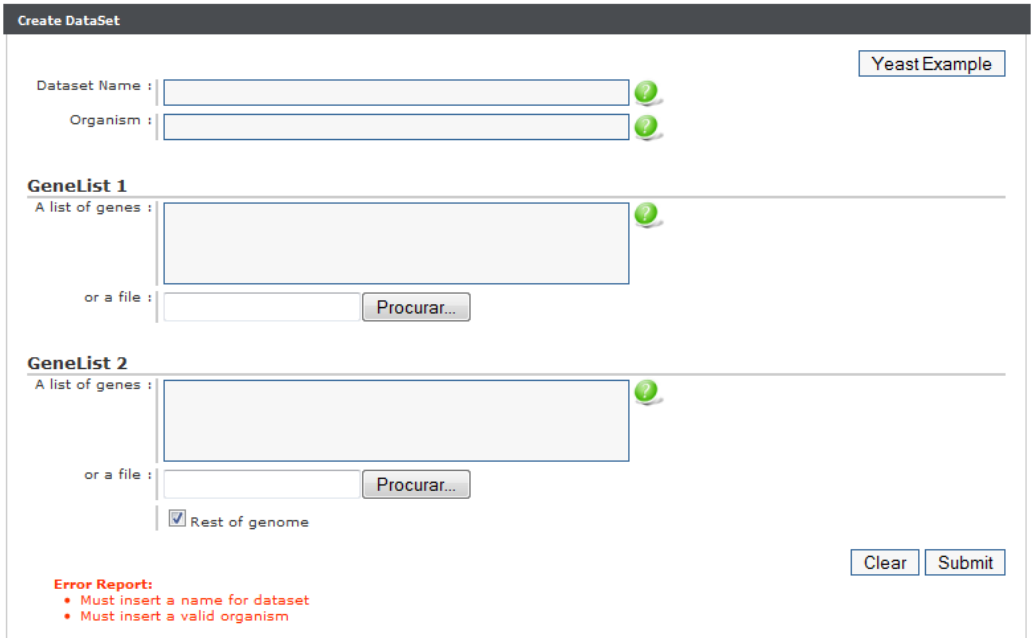



**Figura 4.10 Menu de inserção de um Dataset.**

Como se observa pela Figura 4.10 existem quatro caixas de texto:

- *Dataset Name* é uma forma do utilizador se lembrar dos *Datasets* inseridos
- *Organism* contém a opção de *autocomplete*, quando o utilizador começa a escrever o sistema mostra ao utilizador as opções disponíveis para o nome do organismo
- *Genelist 1*, que não deve conter mais de 300 genes, esta é a lista genes de interesse para o estudo, podem ser um conjunto de genes diferencialmente expressos obtidos de uma experiência de *microarray* ou simplesmente um conjunto de genes que o utilizador deseja testar
- *Genelist 2*, que corresponde à lista completa dos genes a serem analisados. Se o utilizador pretender examinar todos os genes do organismo a maneira mais fácil é seleccionar a opção "*Rest of Genome*".

As duas listas de genes *Genelist 1 e 2* usam vários tipos de identificadores e separadores para a inserção da lista de genes. Estas especificações estão na secção de ajuda da aplicação.

**List of available DataSet's:**

| Dataset Name | Taxonomic Id | Date | Explorer | Remove Dataset | Edit Dataset | Download Dataset |
|---|---|---|---|---|---|---|
| ps1 | 4932 | 2009-03-18 | | | | |
| DatasetExample | 4932 | 2009-03-16 | | | | |
| DatasetExample | 4932 | 2009-03-16 | | | | |
| DatasetExample | 4932 | 2009-03-16 | | | | |
| DatasetExample | 4932 | 2009-03-16 | | | | |
| humancdkn1 | 9606 | 2009-03-10 | | | | |
| humancdkn1 | 9606 | 2009-03-10 | | | | |
| DatasetExample | 4932 | 2009-02-26 | | | | |

**Figura 4.11 Visualização dos datasets inseridos por um utilizador.**

Depois do utilizador inserir um *Dataset*, os dados presentes neste são:

- O nome do *Dataset*
- Um organismo
- Duas listas de genes (uma que é a lista de interesse e a outra todos os genes do estudo)

Todos os dados são validados no GENS.

### 4.3.2. Gene Explorer

O *Gene Explorer* dá ao utilizador uma forma instantânea de acesso a informação que descreve cada gene (Figura 4.12). Estes dados são obtidos a partir de várias fontes de dados públicas, e estas são estruturadas em sete secções distintas:

- **Gene Summary** que contém o nome do gene, nome completo do gene, sinónimos, função que desempenha, localização e a que secção do *Uniprot* pertence (*Swiss-Prot/Trembl*);
- **Gene Ontology** apresentando separadamente o processo biológico, a função molecular e o componente celular;
- **Pathway** onde podemos ver as vias metabólicas presentes no *KEGG Pathway*;
- **Homologies** que é separada em duas secções *KEGG Brite* que contém ortologias e as homologias presentes no *Interprot*, *Prodom*, *Prints* e *TigrFams*;
- **Structure** que contém a estrutura das proteínas que é obtida no *KEGG*;
- **Sequence** a sequência de cada gene que também é obtida no *KEGG*;
- **References** que contém a bibliografia e referências externas para cada gene que obtemos no Uniprot.

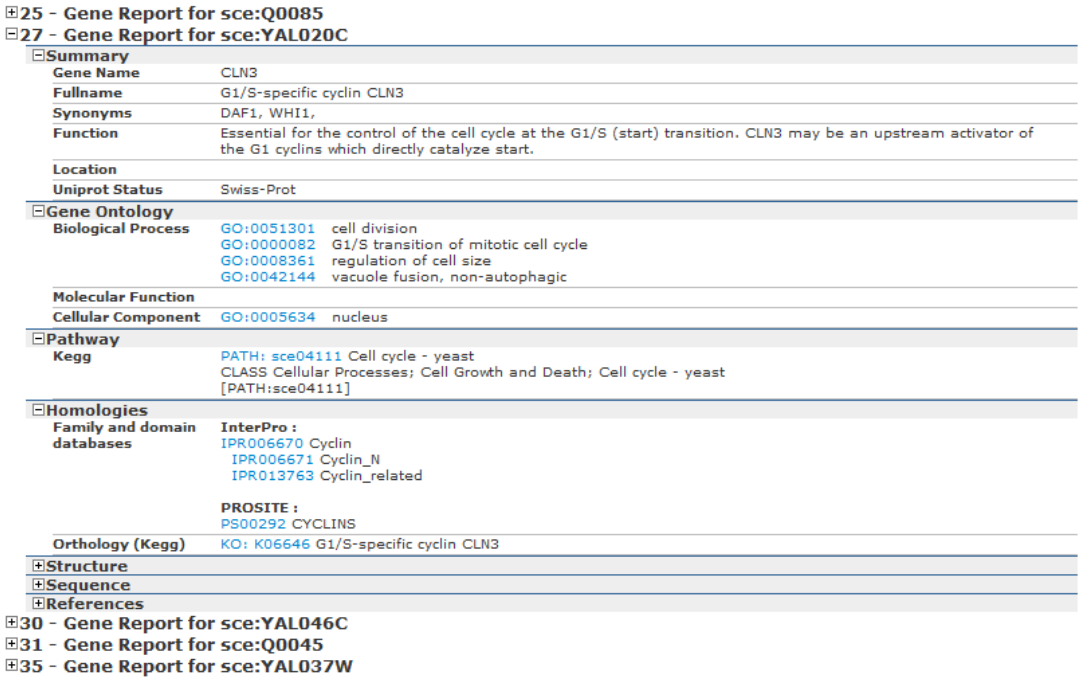



**Figura 4.12 Gridview de visualização de informação no Explorer.**

Do lado esquerdo o utilizador consegue aceder a dois filtros. O primeiro (Figura 4.13), permite ao utilizador seleccionar uma ou mais das sete diferentes secções apresentadas acima (*Summary*,

*Ontology*, *Homology*, etc). Esta funcionalidade dá ao utilizador uma forma simples e rápida de personalizar a informação presente sobre cada gene. A opção foi realizada de forma a permitir ao utilizador personalizar a informação que quer visualizar, utilizadores diferentes podem pretender visualizar informações diferentes.

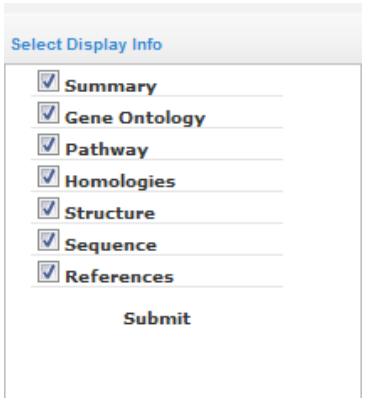



**Figura 4.13 Menu de selecção de informação a mostrar ao utilizador.**

O segundo filtro (*Filter Data*) permite ao utilizador seleccionar informação relativa a uma determinada via metabólica (*Pathway*) ou uma das outras categorias apresentadas (Figura 4.14). Este filtro selecciona todos os genes que estão numa, ou mais do que uma, das classes que se selecciona.

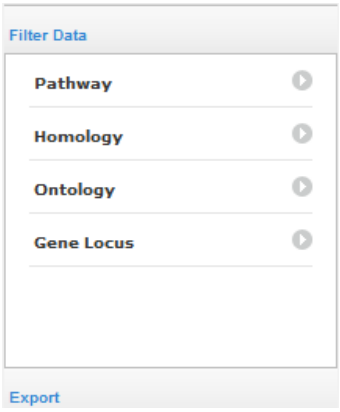



**Figura 4.14 Menu de filtragem de informação a mostrar.**

## 4.3.3. Homology

A homologia entre genes pode referir-se a semelhanças da estrutura, desempenho das mesmas funções, a mesma estrutura em alguma espécie antepassada, entre muitas outras semelhanças.

Existe um grande número de homologias e para cada uma destas há mais que uma fonte de dados. Estas fontes de dados, geralmente, estão concebidas e implementadas em áreas de acesso público.

Devido ao número de tipos de homologia e fontes de dados por tipo de homologia foi feita uma definição clara do problema e é apresentada uma solução.

A solução do problema está definida na Figura 4.15, onde é definido o diagrama com os diferentes tipos de homologia a apresentar ao utilizador. Nesta aplicação estão presentes 4 tipos de homologias: *motifs*, domínios (que podem ser estruturais ou estrutural e funcional), *orthologs* e outros tipos de homologia. Existem ainda várias fontes de dados para cada tipo de homologia como tinha sido referido anteriormente, e devido a esse facto foram seleccionadas a(s) fonte(s) de dados mais relevantes por tipo de homologia.

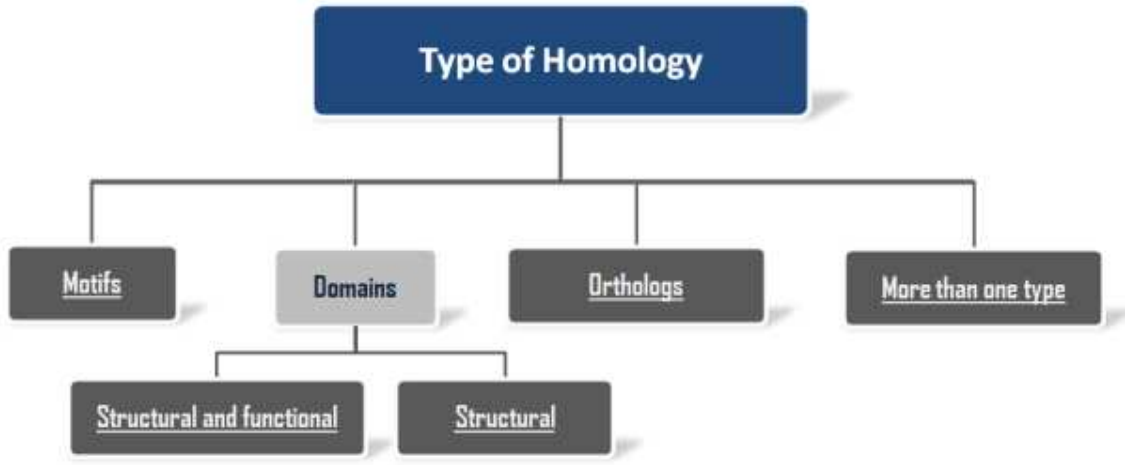



**Figura 4.15 Estrutura da rede de homologias presente no *GeneBrowser*.**

A informação disponibilizada ao utilizador é feita em forma de gráfico, utilizando a livraria construída para esse efeito. No gráfico da Figura 4.16 é apresentada a informação referente a homologia, contendo o p-value, o p-value corrigido, link para a fonte de dados utilizada e ainda a lista de genes por classe de homologia. Com os cálculos do p-value efectuados conseguimos mostrar ao utilizador de uma forma simples e rápida as classes de homologia mais relevantes para o seu estudo.

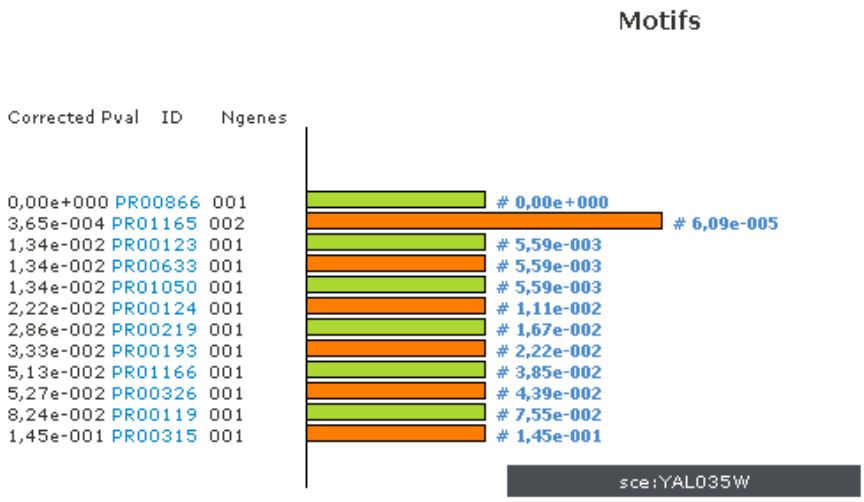



**Figura 4.16 Gráfico representando as classes de homologia.**

Nesta secção o utilizador pode seleccionar a homologia a visualizar. Existem duas formas de selecção da homologia a visualizar: *Select Database* ou *Select Homology Type*.

Na opção *Select Homology Type (*Figura 4.17), o utilizador selecciona o tipo de homologia. A fonte de dados seleccionada para este tipo de homologia é a que se considera mais relevante inicialmente, para esse tipo de homologia.

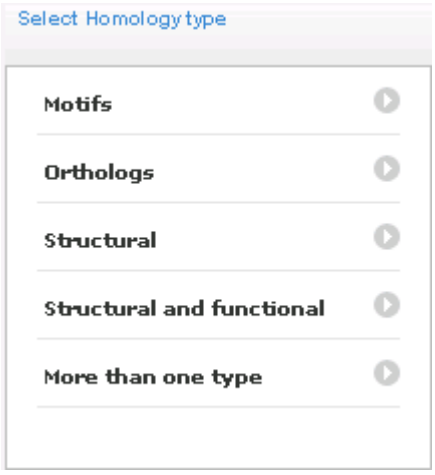



**Figura 4.17 Menu de selecção do tipo de homologia.**

Quando o utilizador selecciona a homologia em *Select Database* (Figura 4.18), o utilizador escolhe a fonte de dados de onde pretende obter as homologias.


**Figura 4.18 Menu de selecção da fonte de dados.**

## 4.3.4. Gene Ontology

A representação das ontologias dos genes é constituída por um vocabulário controlado que descreve o papel dos genes e dos seus produtos. A ontologia é constituída por três grandes classes distintas: processo biológico, função molecular e componentes celulares.

O projecto *Gene Ontology* (GO) (http://www.geneontology.org) classifica os genes numa hierarquia, juntando por função os genes e produtos dos genes. Devido à sua constituição hierárquica, um gene que está numa categoria (Classe) é automaticamente parte da classe pai dessa subclasse e consecutivamente de todos os ramos agregados a este.

O objectivo desta secção do *GeneBrowser* é ajudar o utilizador a obter de uma forma rápida as classes de ontologia com maior relevância estatística. Neste recurso do *GeneBrowser* tem-se dois diferentes tipos de vista: vista em gráfico e vista em árvore.

Para cada vista presente, o utilizador ainda tem a opção de seleccionar uma sub ontologia: processo biológico, função molecular ou componente celular. Existe também a opção do utilizador seleccionar o tipo de vista que pretende do lado esquerdo no *menu*, onde tem a opção de selecção do tipo de vista.

Quando se selecciona a vista em árvore, pode-se navegar pelos vários níveis de ontologia, pela constituição hierárquica das classes ontológicas, por os valores de p-value correspondentes a cada classe de ontologia e pela lista de genes presentes em cada classe (Figura 4.19).

Devido aos elementos da árvore pertencerem a níveis diferentes, não parecia muito lógico apresentar o valor do p-value corrigido, pois este requer uma análise de classes semelhantes. Então a apresentação do p-value corrigido está apenas presente na vista em gráfico onde se pode podar a árvore, analisando assim os seus vários níveis em separado.

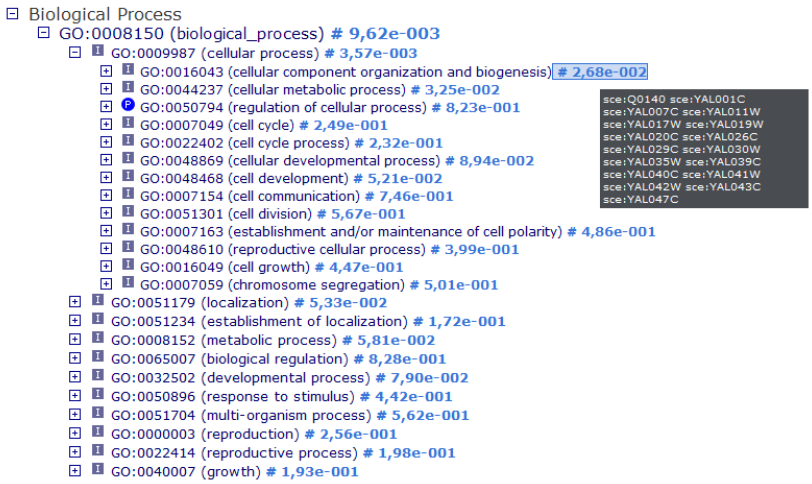



**Figura 4.19 Vista da ontologia em árvore.**

Na vista em gráfico (Figura 4.20) podemos visualizar os vários níveis de ontologia e ainda todas as folhas que contêm genes.

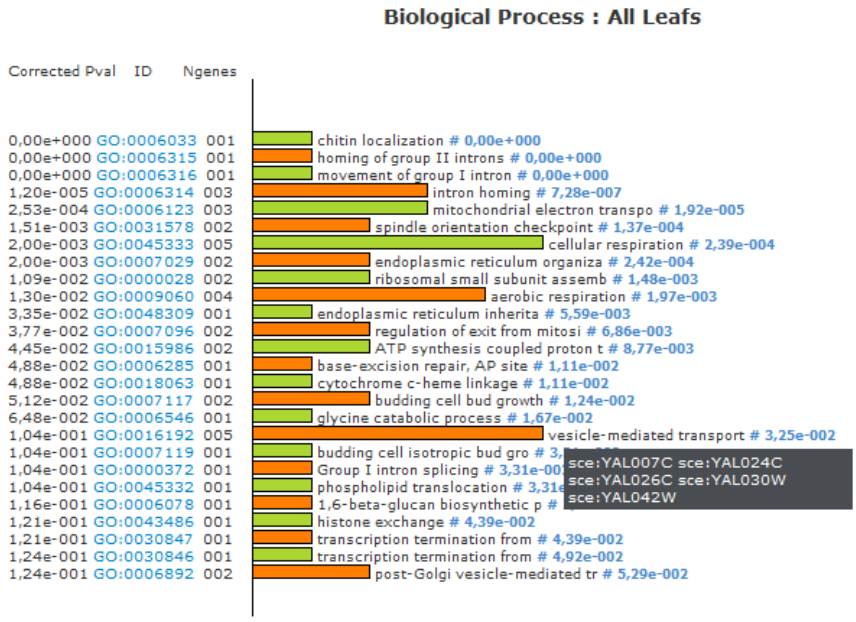



**Figura 4.20 Vista da ontologia em gráfico.**

## 4.3.5. Pathway Explorer

Um *Pathway* (via metabólica) é composto por uma série de reacções bioquímicas que estão ligados por intermédio de reagentes (ou substratos) de uma reacção e os seus produtos. O *GeneBrowser* apenas contém de momento vias metabólicas.

Para facilitar o uso da informação genómica no estudo do processo de atribuição de funções aos genes e aos seus elementos reguladores, é necessária uma perspectiva que permita interligar as sequências do genoma no contexto de vários tipos de informação biológica. As vias de sinalização são o formato lógico de modelação para apresentar essa informação de forma familiar aos biólogos.

A análise presente no *GeneBrowser* contém o enriquecimento de classes funcionais, a fim de identificar as vias de sinalização mais relevantes para o estudo pretendido (Figura 4.21).

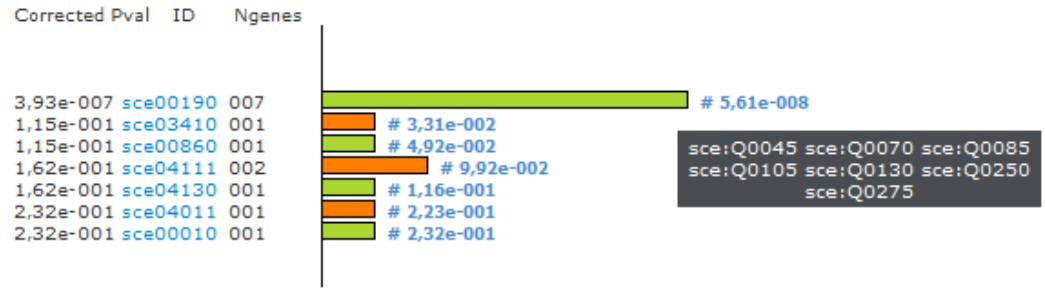



**Figura 4.21 Gráfico contendo informação das vias de sinalização.**

Nesta secção do *software* é apresentado ao utilizador uma forma de observar as mais relevantes vias de sinalização, bem como os genes que estas contêm, além do valor do p-value e p-value corrigido, sendo ainda apresentado um *link* directo para a fonte de dados que contém a anotação dos genes.

## 4.3.6. Gene Expression

Em biologia, grande parte do trabalho de investigação prende-se com as publicações já realizadas sobre o mesmo assunto, e nestas, quando é efectuada uma publicação, geralmente tem de se publicar os dados que levaram a essa publicação. O *Arrayexpress* é um armazém de dados que contém um grande número de experiências de *microarrays* publicadas a partir das quais se podem obter a descrição de cada experiência, os factores experimentais, os valores de factores

experimentais e os correspondentes valores de expressão génica por valores de factores experimentais.

A secção do *GeneBrowser* apresenta como novidade o método de visualização (Figura 4.22). Quando se realiza uma experiência, o resultado é geralmente um conjunto de genes diferencialmente expressos. No *GeneBrowser* apresenta-se ao utilizador uma vista em árvore, onde o primeiro nível da árvore corresponde aos factores experimentais e o segundo aos valores dos factores experimentais. No último nível tem-se as experiências que contêm os genes inseridos pelo utilizador e as experiências em que este esteve presente e os seus valores de expressão, bem como os *links* para o *Arrayexpress*. O conjunto de genes apresentados são os genes diferencialmente expressos em relação ao conjunto de referência.

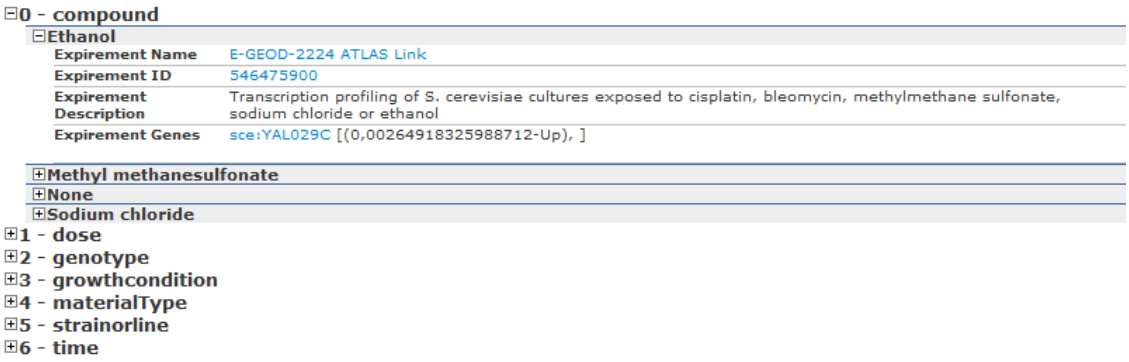



**Figura 4.22 Método de visualização dos dados de expressão génica.**

## 4.3.7. Gene On Locus

Os cromossomas são heterogéneos em estrutura e função. Mas num mesmo cromossoma, genes que estão perto uns dos outros aparecem num mesmo processo biológico, realizando funções biológicas semelhantes e participando nas mesmas interacções.

O que se pretende alcançar com estas características é a de mostrar aos utilizadores uma forma de observar a distribuição dos genes no cromossoma e também ver os cromossomas mais relevantes para o estudo. Nesta secção da aplicação também é utilizado o gráfico desenvolvido previamente (Figura 4.23).

É ainda apresentado ao utilizador a distribuição dos genes no cromossoma, com o cálculo do p-value e p-value corrigido. O sistema também mostra os genes presentes em cada cromossoma e o número de genes do cromossoma. O sistema ainda cria um *link* para *NCBI Mapviewer* onde mostramos a distribuição dos genes no cromossoma.

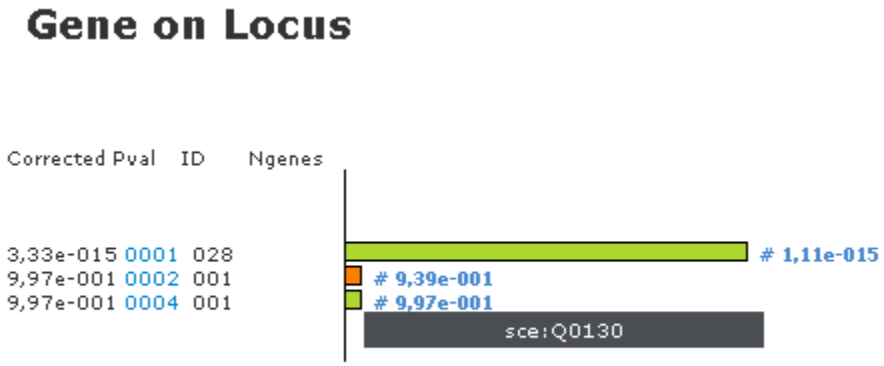



**Figura 4.23 Gráfico representando a distribuição de genes nos cromossomas.**

## 4.3.8. Bibliography

Em biologia, a mais relevante forma de apresentar um trabalho, reside em publicar o trabalho académico numa revista científica. Portanto, a literatura é um relevante caminho para chegar à extracção de conhecimento de um estudo ou questão biológica.

Nesta secção da aplicação, o trabalho concentrou-se em encontrar uma forma de apresentar ao utilizador os artigos mais relevantes para um conjunto de genes.

Para responder a esta questão, teve-se em conta o número de genes, mas também o número total de genes que o artigo apresenta. Num primeiro passo calcula-se o p-value, num segundo passo o corrigido, o terceiro passo multiplicamos o p-value corrigido pelo número de genes e no último passo é feita uma normalização do valor.

Com isso, temos uma maneira de apresentar os cem artigos mais relevantes para o conjunto de genes, que são apresentados pelo seu nível de relevância (Figura 4.24).


**Figura 4.24 Página de apresentação de Bibliography no *GeneBrowser*.**

## 4.4. Sumário

Tendo em conta todo o processo de desenvolvimento e dificuldades encontradas ao longo desse processo, o *GeneBrowser* é uma plataforma sólida. Para a qual foram desenvolvido diferentes componentes, que interligados entre si fornecem ao utilizador uma forma rápida a análise funcional de um conjunto de genes.

# Capítulo 5 - Testes e Validação

Para verificar o devido funcionamento da aplicação foi escolhido um conjunto de genes, indicados na Tabela 5.1. A sua escolha foi feita partindo da importância das vias de sinalização, pois determinadas vias desempenham funções essenciais à célula e ao organismo, segundo uma perspectiva mais global. Por outro lado, os genes que nelas participam tendem a ser melhor estudados (e anotados) e, dependendo da sua função, esses genes podem também estar mais bem conservados evolutivamente – tendo sido sujeitos a conservação (ou sofrido poucas mutações) ou até duplicação e daqui se obtém dados relevantes tanto a nível homológico como ontológico. Além do mais, genes cujos produtos desempenham funções similares tendem a encontrar-se relativamente próximos a nível de *locus*. Assim, estes dados permitem-nos verificar o bom funcionamento da ferramenta, bem como o bom funcionamento da biblioteca estatística e dos dados integrados no GeNS.

**Tabela 5.1 Lista de genes usados para o Dataset de teste.**

| | | | | |
|---|---|---|---|---|
| ACO2 | RPO26 | RFA1 | IDP1 | GUS1 |
| CKS1 | AIM10 | RPC10 | RNH1 | VAS1 |
| RET1 | FDH1 | RFC4 | CHK1 | UGP1 |
| RET1 | PGU1 | GLN4 | FUM1 | FBA1 |
| PET112 | RNH201 | PRS2 | MDH2 | PRS4 |
| SFA1 | TYS1 | PRS2 | RPB11 | PRS4 |
| GRE3 | CDC9 | DED81 | DIA4 | RET1 |
| MDH3 | EXO1 | XYL2 | GND1 | KRS1 |
| CDC27 | RPA190 | BUB3 | RAD24 | POL12 |
| PRS4 | CTA1 | LPD1 | RPO31 | sds |
| ACO1 | THS1 | NAM2 | APC2 | sdsds |
| MCM2 | POL30 | LAT1 | RFC1 | sdsdsd |
| ORC1 | CDC60 | DOC1 | RPB10 | sdsdsd |
| MSK1 | RPC40 | RPA14 | CTA1 | sdsds |
| ZWF1 | NAM2 | PRS2 | WRS1 | sddsd |

## 5.1. Inserção e Validação do Dataset

Na inserção do *Dataset* verificamos que alguns dos genes inseridos não foram validados pelo sistema devido a erros nos identificadores submetidos, demonstrando assim o seu bom funcionamento.

## 5.2. Descrição Geral dos Genes

A mais-valia do Explorer consiste nos processos de filtragens e selecção de informação, referidos aquando da descrição da funcionalidade. Estes permitem comparação e filtragem de genes pertencente às mesmas classes.

Tal como podemos ver na Figura 5.1 a), temos um *dataset* com 69 genes. Quando aplicamos a filtragem por *pathway* seleccionando o *pathway* (*sce03030*) obtemos a Figura 5.1 b) que é a informação filtrada.

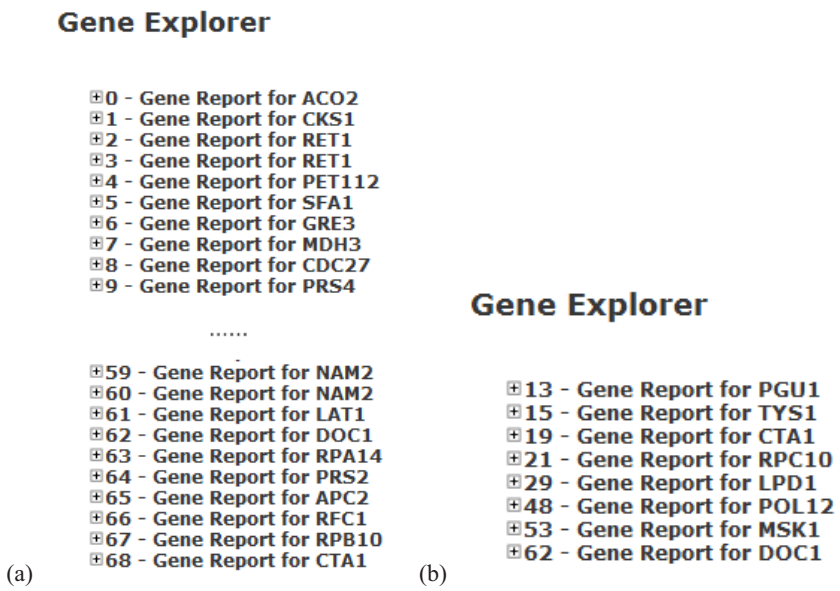



**Figura 5.1 a) Contém uma vista de todos os genes presentes no Dataset. b) Contém um subset de genes que contém o pathway "*sce03030*".**

Deste modo o utilizador tem presente uma forma simples e rápida de filtragem de informação a estudar, conseguindo uma vista simplificada da informação que se torna mais relevante para o seu estudo.

## 5.3. Vias Metabólicas

Viualiza-se na Figura 5.2 uma pré-selecção de 69 genes envolvidos em 18 vias metabólicas, que foi realizada sabendo que haveria diferenças na relevância das vias metabólicas, consoante o papel de cada um na célula: a via de sinalização de replicação do ADN *"sce03030"* teria uma maior relevância do que a da pentose fosfato "*sce00230*", dada a sua importância na replicação celular. Assim, dos 30 genes identificados como pertencentes ao pathway de replicação de ADN, foram introduzidos no sistema apenas 8, aleatoriamente escolhidos; dos mais de 30 genes relativos à via metabólica do ciclo celular, introduziram-se também 8; da via metabólica de metabolismo de purinas, foram seleccionados apenas 10 genes – a mesma metodologia foi utilizada para as vias metabólicas de interconversão de pentose e glucoronato, ciclo de Krebs, ciclo celular, pentose fostato, biossíntese de valina, leucina e isoleucina, metabolismo de triptofano, glicólise/gluconeogénese, metabolismo do glutamato e as restantes vias metabólicas.

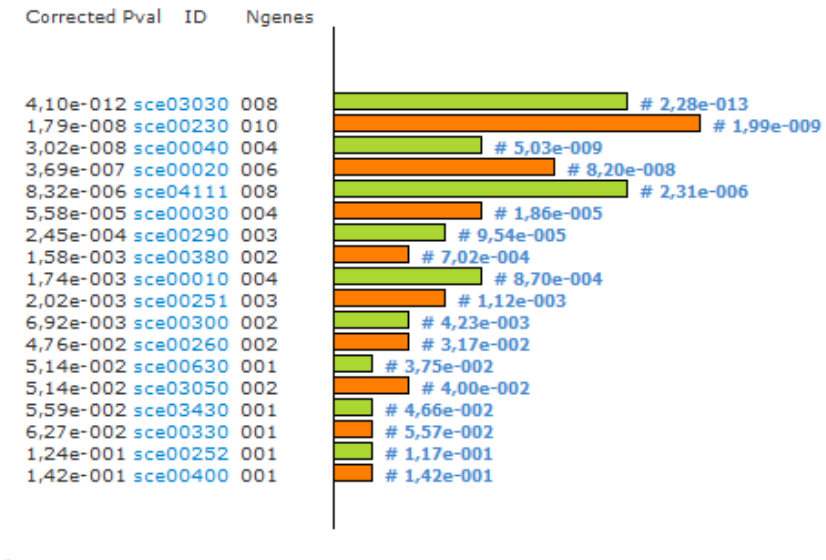



**Figura 5.2 Gráfico representando a distribuição dos genes nos Pathways e relativo p-value.**

Após a construção do *dataset*, a aplicação indicou que os *pathways* mais relevantes seriam: replicação de ADN, seguido de síntese de purinas, interconversão de pentose e glucoronato, ciclo

de *Krebs* e o ciclo celular, para citar apenas os 5 primeiros resultados. O que os resultados demonstram é que a aplicação calculou um valor de p-value corrigido para o *pathway* da replicação de ADN inferior ao valor para o pathway do ciclo de *Krebs*; biologicamente, a replicação de ADN é um evento extremamente importante, ainda que ocorrendo menos vezes do que o ciclo de Krebs, e que é crucial para o correcto funcionamento das células filhas – qualquer erro no processo de replicação de ácidos nucleicos pode originar células com defeitos enzimáticos, cujas consequências dependerão do tipo de erro que tenha surgido inicialmente. Porém, sem o ciclo de *Krebs*, a célula fica sem produtos a partir dos quais obter energia (o processo não será directo, pois o ciclo *Krebs* está envolvido num processo metabólico de obtenção de energia que é bastante mais complexo e envolve outros *pathways*) e como tal é igualmente um processo extremamente relevante, que pode ser confirmado pelo valor de p-value corrigido, que coloca o Pathway nos topo dos resultados.

## 5.4. Homologias

Dois organismos que partilhem características associadas a um ancestral comum dizem-se homólogos. O mesmo conceito pode ser aplicado a estruturas moleculares, sendo bastante utilizado no estabelecimento de relações filogenéticas entre espécies.

A nível proteico, descobriu-se que existem determinadas zonas das sequências de aminoácidos que são mantidas quase inalteradas entre espécies próximas, *i.e.*, sequências proteicas que sofreram poucas (ou nenhumas) alterações durante a história evolutiva dessas espécies. Essas zonas denominam-se locais conservados (*conserved site*). No *dataset* de testes que usámos, havíamos introduzido alguns genes associados ao mesmo *pathway* e cujos produtos desempenhavam funções semelhantes. Sabendo que genes cujos produtos participam nas mesmas vias de sinalização ou cujos produtos desempenham funções semelhantes, tendem a partilhar características homólogas. Observa-se na Figura 5.3 os valores de p-value e a distribuição dos genes pelas diferentes classes de homologia.

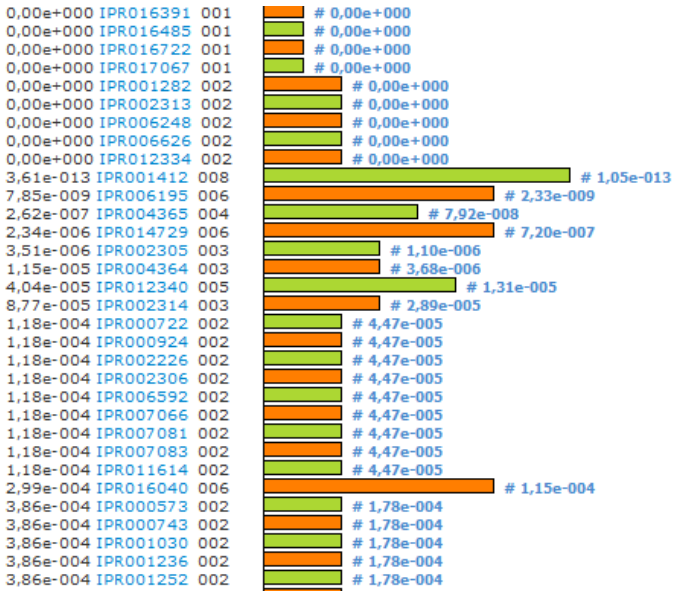



**Figura 5.3 P-value e distribuição dos genes por classes de homólogos.**

## 5.5. Descrição Ontológica

Nos organismos unicelulares, a sua única célula contém todas as funções necessárias à sobrevivência e reprodução do organismo. Os organismos multicelulares porém, apresentam várias células, agrupadas em tecidos e órgãos, em que cada órgão desempenha uma função. Assim, uma célula enquanto elemento de um organismo multicelular, está dependente das outras do mesmo organismo para que possa sobreviver e funcionar correctamente. Não pretendendo diferenciar as prioridades intrínsecas a cada célula, e havendo escolhido alguns *pathways* para testar a estruturação ontológica da aplicação, sobretudo *pathways* que se sabia serem importantes e praticamente ubíquos – como é o caso de ligações proteicas, ou de ligações de amino-ácidos a moléculas tARN – era de esperar que a aplicação mostrasse que os *pathways* mais relevantes fossem aqueles cujos genes teriam ontologias associadas a processos metabólicos de obtenção de energia, a degradação/construção de compostos ou à replicação celular (ou mecanismos de correcção de erros) pois são estas as funções que a célula, ainda que dependente do resto do organismo, necessita de desempenhar.

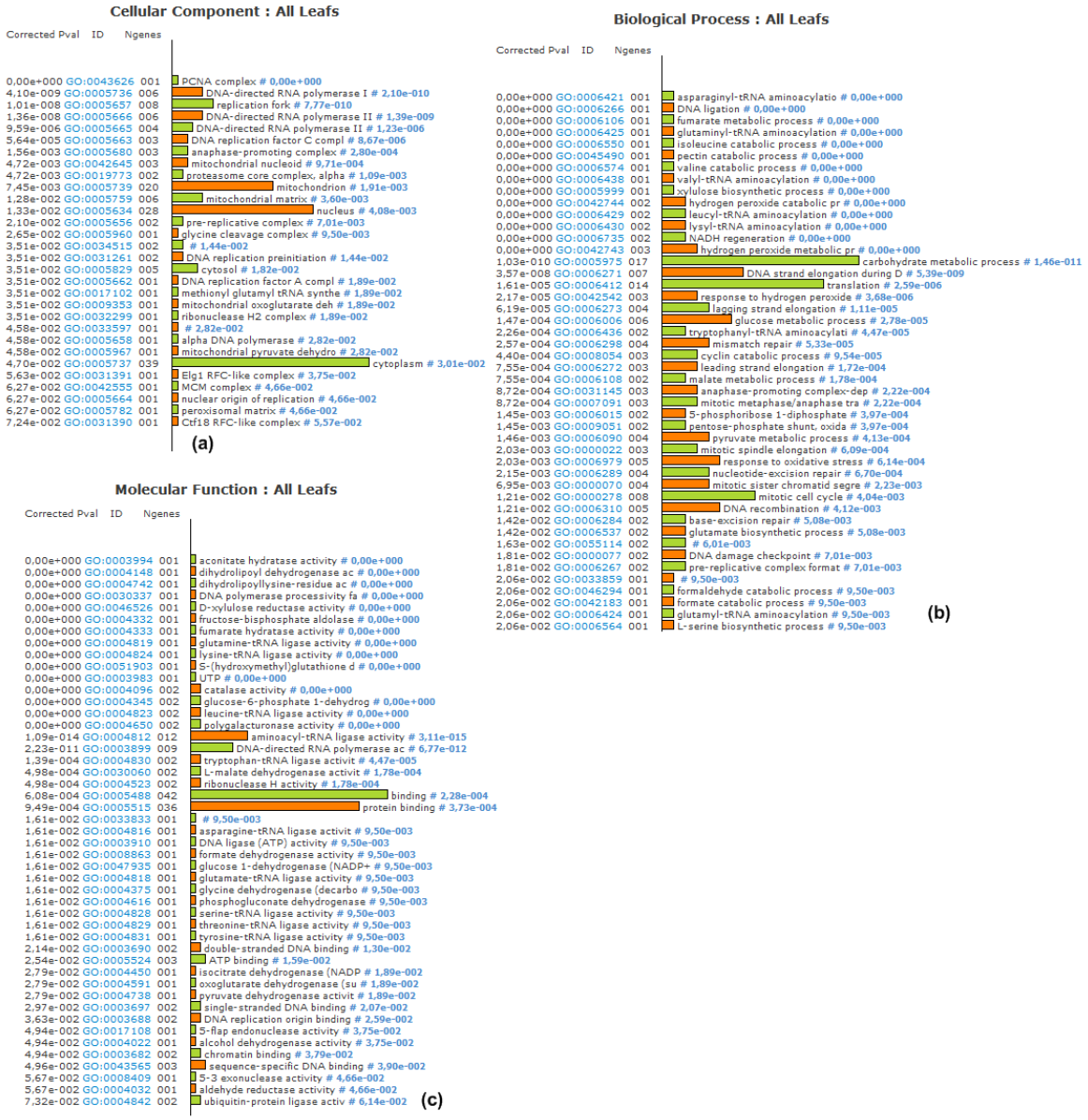



**Figura 5.4 Gráficos das ontologias presentes no Dataset testado: Componente Celular, b) Processo Biológico e c) Função molecular.**

Os resultados obtidos na ontologia para o *dataset* de testes, foram muito próximo dos esperados como pode-se ver na Figura 5.4. As classes ontológicas mais relevantes estão associadas a processos metabólicos de obtenção de energia, a degradação/construção de compostos ou à replicação celular.

## 5.6. Localização dos Genes no Cromossoma

Foram seleccionados genes de três cromossomas I, II e IV. Os genes desses cromossomas foram seleccionados de modo a que o cromossoma I tivesse maior significância que os outros dois. Assim sendo, seria de esperar da análise efectuada pelo *GeneBrowser* que o cromossoma I fosse mais significante que os outros dois, como se confirma na Figura 5.5

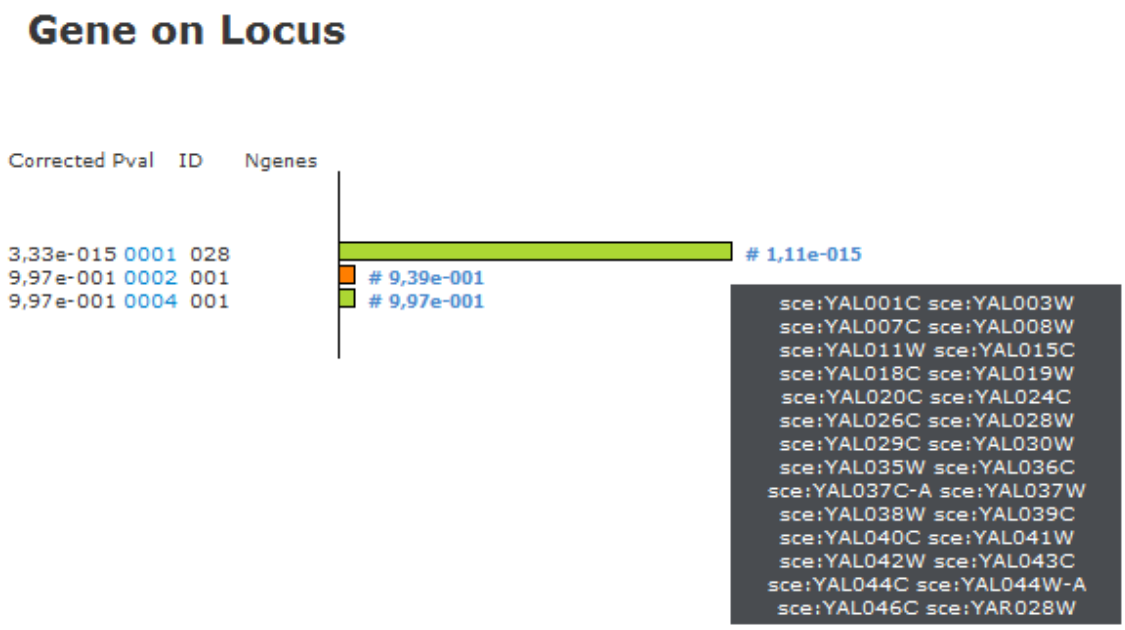


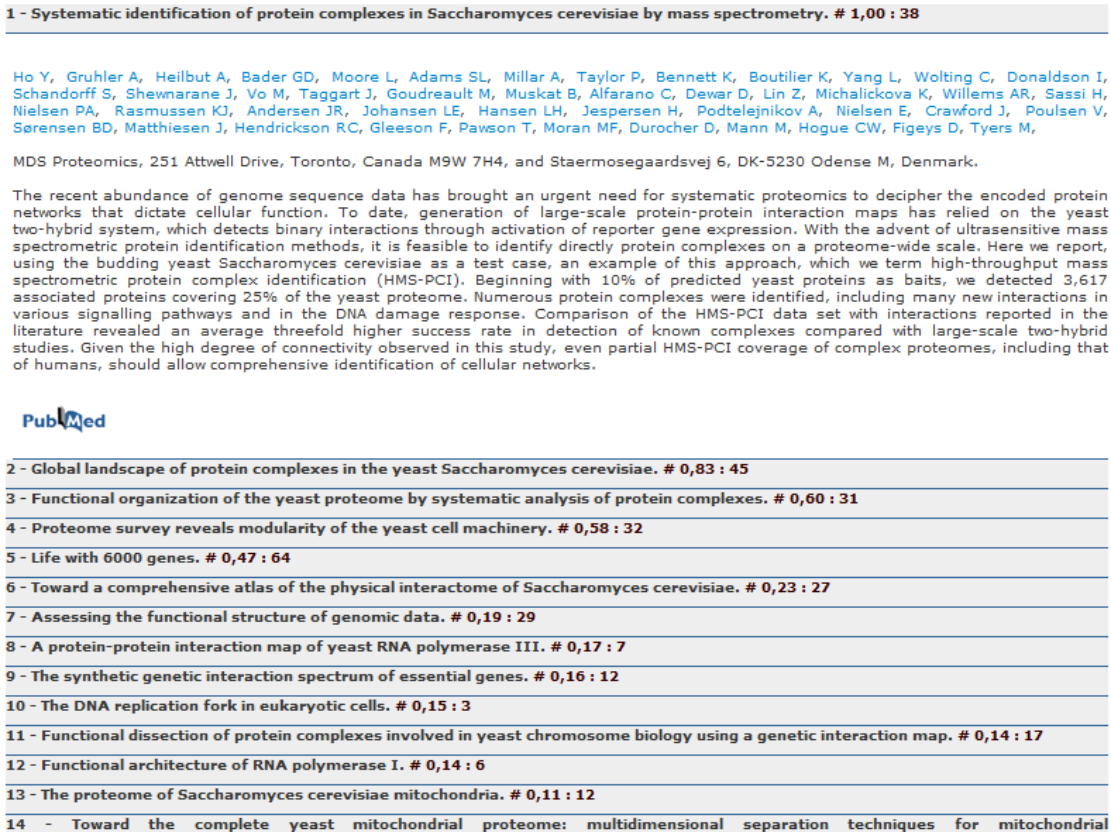

**Figura 5.5 P-value e distribuição dos genes nos cromossomas.**

## 5.7. Bibliografia

Os artigos mais relevantes descrevem a maquinaria celular do organismo, porque uma vez feita essa descrição os estudos subsequentes podem basear-se nesses artigos.

Artigos como o "The DNA replication fork in eukaryotic cells. # 0,15 : 3" faz sentido estarem presentes apesar do baixo número de genes porque descrevem o mecanismo que é de extrema importância para o ciclo celular e para a compreensão das alterações genéticas que podem surgir apenas de não ser específico para o organismo. Por outro lado artigos como: "*A protein-protein interaction map of yeast RNA polymerase III. # 0,17 : 7*" são específicos do organismo, mesmo ainda que o processo estudado seja ubíquo, ainda está com um *ranking* superior.

Na Figura 5.6 são apresentados os artigos mais relevantes e o *ranking* atribuído a estes pelo *GeneBrowser*.



**1 - Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. # 1,00 : 38**

Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, Adams SL, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems AR, Sassi H, Nielsen PA, Rasmussen KJ, Andersen JR, Johansen LE, Hansen LH, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, Sørensen BD, Matthiesen J, Hendrickson RC, Gleeson F, Pawson T, Moran MF, Durocher D, Mann M, Hogue CW, Figeys D, Tyers M,

MDS Proteomics, 251 Attwell Drive, Toronto, Canada M9W 7H4, and Staermosegaardsvej 6, DK-5230 Odense M, Denmark.

The recent abundance of genome sequence data has brought an urgent need for systematic proteomics to decipher the encoded protein networks that dictate cellular function. To date, generation of large-scale protein-protein interaction maps has relied on the yeast two-hybrid system, which detects binary interactions through activation of reporter gene expression. With the advent of ultrasensitive mass spectrometric protein identification methods, it is feasible to identify directly protein complexes on a proteome-wide scale. Here we report, using the budding yeast Saccharomyces cerevisiae as a test case, an example of this approach, which we term high-throughput mass spectrometric protein complex identification (HMS-PCI). Beginning with 10% of predicted yeast proteins as baits, we detected 3,617 associated proteins covering 25% of the yeast proteome. Numerous protein complexes were identified, including many new interactions in various signalling pathways and in the DNA damage response. Comparison of the HMS-PCI data set with interactions reported in the literature revealed an average threefold higher success rate in detection of known complexes compared with large-scale two-hybrid studies. Given the high degree of connectivity observed in this study, even partial HMS-PCI coverage of complex proteomes, including that of humans, should allow comprehensive identification of cellular networks.

PubMed

**2 - Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. # 0,83 : 45**

**3 - Functional organization of the yeast proteome by systematic analysis of protein complexes. # 0,60 : 31**

**4 - Proteome survey reveals modularity of the yeast cell machinery. # 0,58 : 32**

**5 - Life with 6000 genes. # 0,47 : 64**

**6 - Toward a comprehensive atlas of the physical interactome of Saccharomyces cerevisiae. # 0,23 : 27**

**7 - Assessing the functional structure of genomic data. # 0,19 : 29**

**8 - A protein-protein interaction map of yeast RNA polymerase III. # 0,17 : 7**

**9 - The synthetic genetic interaction spectrum of essential genes. # 0,16 : 12**

**10 - The DNA replication fork in eukaryotic cells. # 0,15 : 3**

**11 - Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. # 0,14 : 17**

**12 - Functional architecture of RNA polymerase I. # 0,14 : 6**

**13 - The proteome of Saccharomyces cerevisiae mitochondria. # 0,11 : 12**

**14 - Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial**

**Figura 5.6 Artigos e ranking calculado pelo *GeneBrowser*.**

Para efectuar um teste no método de *ranking* da bibliografia foi seleccionado um outro *dataset* que continha 28 Genes no Cromossoma I, um gene no Cromossoma II e um gene no Cromossoma IV.

Como se pode ver na Figura 5.7 o primeiro artigo que aparece é um artigo referente ao Cromossoma I de *Saccharomyces Cerevisiae*.

Tendo em conta os resultados obtidos com os dois testes efectuados, pode-se verificar o bom funcionamento do método de *ranking* aplicado.

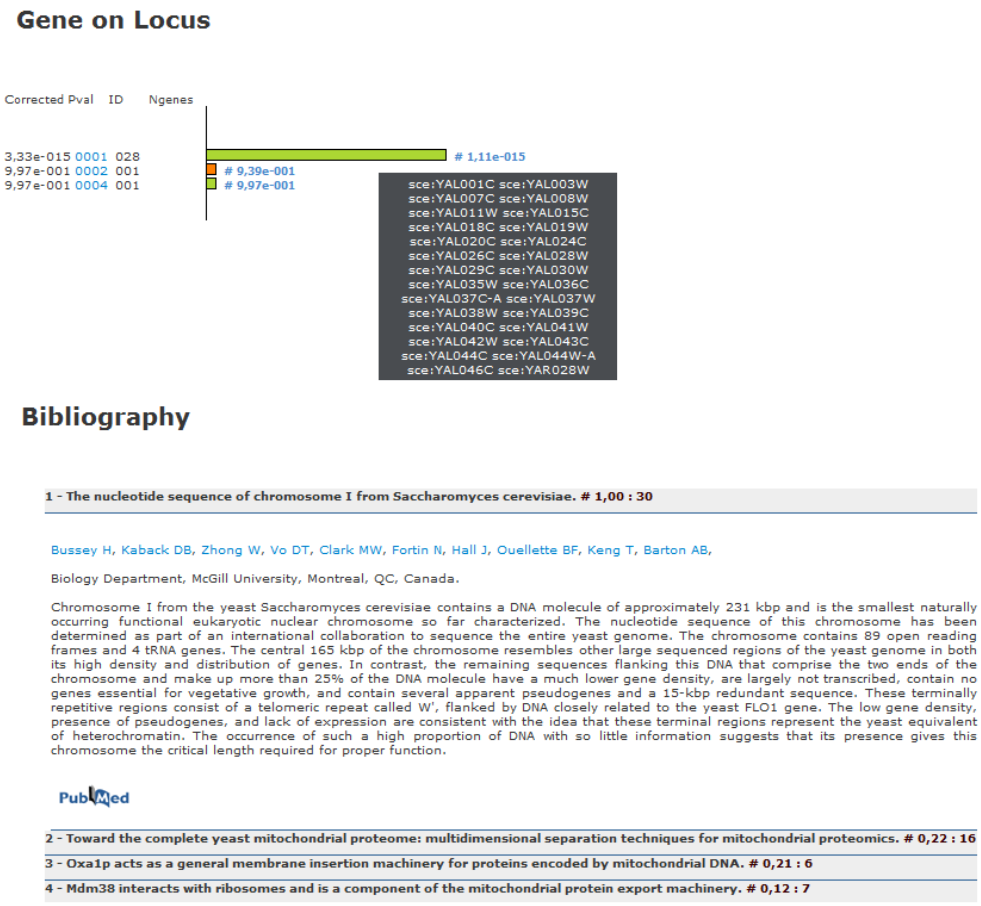



**Figura 5.7 Figura contendo a bibliografia.**

## 5.8. Sumário

Neste capítulo foram apresentados os testes efectuados que demonstram a viabilidade e validade da ferramenta. A ferramenta demonstra ter um bom desempenho, quer a nível de resultados obtidos, quer a nível de desempenho devido ao processamento que necessita.

# Capítulo 6 - Conclusão e Trabalho Futuro

Nesta secção serão analisados os objectivos atingidos com a plataforma do *GeneBrowser*, os conhecimentos aprofundados com a sua elaboração e será feito um balanço geral do estado da plataforma.

## 6.1. Objectivos

Os principais objectivos propostos para este documento foram claramente atingidos:

**Base Tecnológica –** definição de uma arquitectura tecnológica baseada no alto desempenho com processamento intensivo de informação e métodos de cálculo estatístico factorial, que fornecem um forte suporte para a sua extensibilidade;

**Base Estrutural –** através da análise profunda de todas as tarefas e de todo o contexto envolvente, definiu-se uma base estrutural forte, baseada nas boas regras de modulação e estabelecendo uma forte ligação entre conceitos, abrindo as portas para futuras evoluções;

**Base Interactiva –** conhecendo os seus utilizadores e baseado em conhecimentos maiores da usabilidade, definiu-se uma base interactiva extremamente organizada e de fácil utilização, dinamizando a realização de todas as tarefas antes complexas;

## 6.2. Aprendizagem

Ao longo do desenvolvimento desta complexa plataforma, foi necessário obter conhecimentos profundos acerca de todo o processo de desenvolvimento e publicação de aplicações *Web*. A nível de linguagens de programação foram obtidos ou aprofundados conhecimentos em HTML, CSS, XML, ASP.NET, ASP.NET AJAX, C#, *JavaScript*, JSON e SQL. Na utilização de ferramentas de desenvolvimento constaram o *Microsoft SQL Server Management Studio* e o *Microsoft Visual Studio*. Para o SGBD teve de se explorar profundamente o *Microsoft SQL Server*. Na publicação da plataforma desenvolvida, exigiu-se a exploração e configuração do sistema operativo *Microsoft Windows Server* e da aplicação IIS (*Internet Information Services*). Para além de toda esta panóplia

de conhecimentos tecnológicos aprofundados ou adquiridos, houve também uma aprendizagem de interacção social impossível de ignorar, resultado do constante contacto com todos os elementos pertencentes ao grupo de bioinformática.

## 6.3. Balanço Geral

Para fazer um balanço geral de todo o cenário que envolve a plataforma GeneBrowser, foi elaborada uma análise SWOT (*Strengths, Weaknesses, Opportunities, Threats*), proporcionando um resumo estruturado das características que favorecem ou não a ferramenta a nível interno e externo [53].


**Figura 6.1 Diagrama de análise de SWOT.**

## 6.4. Sugestões de Trabalho Futuro

Embora, neste momento o *GeneBrowser* já facilite a tarefa de interpretação biológica, muitas outras funcionalidades lhe podem ser acrescentadas de modo a melhorar as potencialidades actuais. Com o arranque desta nova plataforma, existe um grande conjunto de tarefas que podem ser exploradas e elaboradas, tanto a nível de novas funcionalidades como na utilização de novas técnicas para dinamizar o funcionamento interno da ferramenta.

Das principais conclusões e dificuldades deste trabalho de investigação, são sugeridos alguns tópicos de investigação para melhorar a partir dos resultados obtidos de uma experiência de expressão génica:

- Utilização de outros métodos estatísticos;
- Inserção de uma funcionalidade que relacione os genes com os seus factores de transcrição;
- Uma nova funcionalidade que relacione os genes com doenças e drogas;
- Permitir ao utilizador a selecção do método estatístico a utilizar.

Outra funcionalidade que pode ser explorada consiste na abstracção do SGBD e a utilização de uma camada de abstracção do SGBD que é sempre uma boa técnica a utilizar, não ficando dependente do gestor de base de dados utilizado. De igual modo, será interessante estudar o modelo de programação de persistência de entidades, reunindo as vantagens de utilizar um modelo orientado a objectos na construção de uma aplicação, com o desempenho e confidencialidade das bases de dados actuais. Desta forma, será feito um mapeamento das tabelas da base de dados para a aplicação, deixando de haver acesso directo ao SGBD, utilizando-se apenas os objectos mapeados. Esta metodologia simplifica e dinamiza bastante a construção de funcionalidades.

# Bibliografia


1. Lopes, F.C., et al., *Relatório de Projecto de Final de Curso*. 2006, Universidade de Aveiro: Aveiro.

2. Khatri, P., et al., *Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments.* Nucleic Acids Res, 2004. **32**(Web Server issue).

3. Al-Shahrour, F., et al., *FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments.* Nucleic Acids Res, 2007. **35**(Web Server issue).

4. Zeeberg, B.R., et al., *GoMiner: a resource for biological interpretation of genomic and proteomic data.* Genome Biol, 2003. **4**(4).

5. Dahlquist, K.D., et al., *GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways.* Nat Genet, 2002. **31**(1): p. 19-20.

6. Mlecnik, B., et al., *PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W633-7.

7. Arrais, J., et al. *GeneBrowser: an approach for integration and functional classification of genomic data*. 2007.

8. GOConsortium. *the Gene Ontology*. 1999 19-03-2008 [cited; Available from: http://www.geneontology.org/.

9. Ogata, H., et al., *Computation with the KEGG pathway database.* Biosystems, 1998. **47**(1-2): p. 119-28.

10. Sandler, M.P., *PubMed Central: the JNM perspective.* J Nucl Med, 2000. **41**(7): p. 1123-4.

11. Maglott, D., et al., *Entrez Gene: gene-centered information at NCBI.* Nucleic Acids Res, 2007. **35**(Database issue): p. D26-31.

12. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue).

13. Apweiler, R., et al., *The InterPro database, an integrated documentation resource for protein families, domains and functional sites.* Nucleic Acids Res, 2001. **29**(1).

14. Sonnhammer, E.L., S.R. Eddy, and R. Durbin, *Pfam: a comprehensive database of protein domain families based on seed alignments.* Proteins, 1997. **28**(3): p. 405-20.

15. Rebhan, M., et al., *GeneCards: integrating information about genes, proteins and diseases.* Trends Genet, 1997. **13**(4): p. 163.

16. Zeeberg, B.R., et al., *High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID).* BMC Bioinformatics, 2005. **6**.

17. Baldi, P. and S. Brunak, *Bioinformatics: the machine learning approach*. 2001: MIT press.

18. Mount, D.W., *Bioinformatics: sequence and genome analysis*. 2004: CSHL Press.

19. Galperin, M.Y., *The Molecular Biology Database Collection: 2006 update.* Nucleic Acids Res, 2006. **34**(Database issue): p. D3-5.

20. Galperin, M.Y. and G.R. Cochrane, *Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009.* Nucleic Acids Res, 2009. **37**(Database issue): p. D1-4.

21. EBI. *European Bioinformatics Institute (EBI)*. 2009 [cited 2009; Available from: http://www.ebi.ac.uk/.

22. WTSI/EBI. *Ensembl*. 2009 [cited 2009 17-04-2009]; Available from: http://www.ensembl.org/.

23. Shah, S.P., et al., *Atlas - a data warehouse for integrative bioinformatics.* BMC Bioinformatics, 2005. **6**: p. 34.

24. Yeats, C., et al., *Gene3D: comprehensive structural and functional annotation of genomes.* Nucleic Acids Res, 2008. **36**(Database issue): p. D414-8.

25. Thomas, P.D., et al., *PANTHER: a library of protein families and subfamilies indexed by function.* Genome Res, 2003. **13**(9): p. 2129-41.

26. Wu, C.H., et al., *PIRSF: family classification system at the Protein Information Resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D112-4.

27. Attwood, T.K., et al., *PRINTS--a database of protein motif fingerprints.* Nucleic Acids Res, 1994. **22**(17).

28. Corpet, F., J. Gouzy, and D. Kahn, *The ProDom database of protein domain families.* Nucleic Acids Res, 1998. **26**(1): p. 323-6.

29. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Res, 2006. **34**(Database issue): p. D227-30.

30. Schultz, J., et al., *SMART, a simple modular architecture research tool: identification of signaling domains.* Proc Natl Acad Sci U S A, 1998. **95**(11): p. 5857-64.

31. Hunter, S., et al., *InterPro: the integrative protein signature database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D211-5.

32. Mulder, N. and R. Apweiler, *InterPro and InterProScan: tools for protein sequence classification and comparison.* Methods Mol Biol, 2007. **396**: p. 59-70.

33. Quevillon, E., et al., *InterProScan: protein domains identifier.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W116-20.

34. Parkinson, H., et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): p. D747-50.

35. Parkinson, H., et al., *ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression.* Nucleic Acids Res, 2009. **37**(Database issue): p. D868-72.

36. Wheeler, D.L., et al., *Database resources of the national center for biotechnology information.* Nucleic acids research, 2007. **35**(Database issue): p. D5.

37. Hamosh, A., et al., *Online Mendelian Inheritance in Man (OMIM).* Hum Mutat, 2000. **15**(1): p. 57-61.

38. Kanehisa, M., et al., *From genomics to chemical genomics: new developments in KEGG.* Nucleic Acids Res, 2006. **34**(Database issue): p. D354-7.

39. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2008. **36**(Database issue): p. D480-4.

40. Ogata, H., et al., *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): p. 29-34.

41. Haft, D.H., J.D. Selengut, and O. White, *The TIGRFAMs database of protein families.* Nucleic Acids Res, 2003. **31**(1): p. 371-3.

42. Galperin, M.Y., *The Molecular Biology Database Collection: 2008 update.* Nucleic Acids Res, 2007.

43. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2007. **35**(Database issue): p. D21-5.

44. Kanehisa, M., et al., *KEGG for linking genomes to life and the environment.* Nucleic Acids Res, 2007.

45. Detours, V., et al., *Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets.* FEBS Lett, 2003. **546**(1): p. 98-102.

46. Achard, F., G. Vaysseix, and E. Barillot, *XML, bioinformatics and data integration.* Bioinformatics, 2001. **17**(2): p. 115-25.

47. Lee, T.J., et al., *BioWarehouse: a bioinformatics database warehouse toolkit.* BMC Bioinformatics, 2006. **7**: p. 170.

48. Kuntzer, J., et al., *BNDB - the Biochemical Network Database.* BMC Bioinformatics, 2007. **8**: p. 367.

49. Birkland, A. and G. Yona, *BIOZON: a hub of heterogeneous biological data.* Nucleic Acids Res, 2006. **34**(Database issue): p. D235-42.

50. Al-Shahrour, F., R. Diaz-Uriarte, and J. Dopazo, *FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes.* Bioinformatics, 2004. **20**(4).

51. Zhang, B., et al., *GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies.* BMC Bioinformatics, 2004. **5**: p. 16.

52. Cases, I., et al., *CARGO: a web portal to integrate customized biological information.* Nucleic Acids Res, 2007. **35**(Web Server issue): p. W16-20.

53. Oliveira, J.L., et al. *DiseaseCard: A Web-Based Tool for the Collaborative Integration of Genetic and Medical Information*. 2004: Springer.

54. Al-Shahrour, F., et al., *Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments.* Nucleic Acids Res, 2008. **36**(Web Server issue).

55. Al-Shahrour, F., et al., *BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments.* Nucleic Acids Res, 2006. **34**(Web Server issue).

56. Al-Shahrour, F., et al., *BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments.* Nucleic Acids Res, 2005. **33**(Web Server issue).

57. Primo, A. *O aspecto relacional das interações na Web 2.0*. 2006.

58. Garrett, J.J., *Ajax: A New Approach to Web Applications*. 2005. 1-5.

59. Goodman, D., M. Morrison, and Books24x7 Inc., *JavaScript bible, sixth edition*. 2007, Wiley Pub.: Indianapolis, Ind.

60. Easttom, C. and Books24x7 Inc., *Advanced JavaScript, third edition*. 2008, Wordware Pub.: Plano, Tex.

61. Zakas, N.C., *Professional JavaScript for Web developers*. Wrox professional guides. 2005, Hoboken, N.J.: Wiley Pub.

62. McComb, G., *Extensible Markup Language (XML) 1.0 specifications*. 1999, San Jose, CA: toExcel. p.

63. Shanmugasundaram, J., et al. *Relational Databases for Querying XML Documents: Limitations and Opportunities*. in *Proceedings of the 25th VLDB Conference*. 1999. Edinburgh.

64. Croft, J., et al., *Pro CSS techniques*. 2006, Apress: Berkeley, Calif.

65. Cerami, E., *Web services essentials*. 1st ed. 2002, Beijing ; Sebastopol, CA: O'Reilly. xiii, 288 p.

66. Sirin, E., J. Hendler, and B. Parsia, *Semi-automatic Composition of Web Services using Semantic Descriptions.*

67. Draghici, S., *Data analysis tools for DNA microarrays*. 2003: CRC Press.

68. Gardner, M.J. and D.G. Altman, *Confidence intervals rather than P values: estimation rather than hypothesis testing.* British Medical Journal, 1986. **292**(6522): p. 746.